UNIVERSIDADE FEDERAL DO PARANÁ

ALANA RENATA RIBEIRO

MÉTODOS NUMÉRICOS APLICADOS À DETECÇÃO DE ANOMALIAS EM DADOS DE VAZÃO

CURITIBA
2014

ALANA RENATA RIBEIRO

MÉTODOS NUMÉRICOS APLICADOS À DETECÇÃO DE ANOMALIAS EM DADOS DE VAZÃO

Dissertação apresentada ao Curso de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação Matemática, do Departamento de Matemática, Setor de Ciências Exatas e do Departamento de Construção Civil, Setor de Tecnologia, Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Mestre em Ciências.

Orientadora: Profª. Drª. Deise Maria Bertholdi Costa
Coorientador: Dr. Eduardo Alvim Leite

CURITIBA
2014

R484m Ribeiro, Alana Renata
Métodos numéricos aplicados à detecção de anomalias em dados de vazão / Alana Renata Ribeiro. – Curitiba, 2014.
95f. : il. color. ; 30 cm.

Dissertação (mestrado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-graduação em Métodos Numéricos em Engenharia, 2014.

Orientadora: Deise Maria Bertholdi Costa -- Coorientador: Eduardo Alvim Leite.
Bibliografia: p. 91-93.

1. Vazante. 2. Hidrologia - Modelos. 3. Redes neurais I. Universidade Federal do Paraná. II. Costa, Deise Maria Bertholdi III. Leite, Eduardo Alvim IV. Título.

CDD: 551.489

TERMO DE APROVAÇÃO

ALANA RENATA RIBEIRO

MÉTODOS NUMÉRICOS APLICADOS À DETECÇÃO DE ANOMALIAS EM DADOS DE VAZÃO.

Dissertação aprovada como requisito parcial para obtenção do grau de mestre no Programa de Pós-Graduação em Métodos Numéricos em Engenharia, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Prof.ª Dr.ª Deise Maria Bertholdi Costa
Orientadora – Membro do PPGMNE/UFPR

Prof. Dr. Paulo Henrique Siqueira
Membro do PPGMNE/UFPR

Prof. Dr. Eduardo Felga Gobbi
Membro do Departamento de Engenharia Ambiental da UFPR.

Curitiba, 07 de fevereiro 2014.

AGRADECIMENTOS

A Deus que tanto tem acertado em minha vida, guiado as minhas escolhas, perdoado os meus erros e atendido os meus pedidos.

À minha mãe Iraci, em quem eu me espelho, meu exemplo a seguir, sempre tão atenciosa, prestativa, disposta a ajudar em tudo, pelo seu amor incondicional, carinho, ternura, e a cima de tudo por seu incentivo e conselhos sempre tão bem vindos em toda a minha vida acadêmica e pessoal. Ao meu pai Rogério, que mesmo de longe, acompanha meu desenvolvimento com todo orgulho que um pai pode ter por uma filha, pelas palavras de incentivo nos momentos de dúvidas, paciência nos momentos difíceis, por me ensinar a ser independente, e por toda a personalidade que herdei. À minha irmã Fabiane, aquela em quem sempre posso confiar, que me apoia, mas também me corrige, obrigada pela amizade verdadeira, e pela segurança que me transmite. Obrigada por ter me dado sobrinhos lindos, o Pedro e a Isabelle, que me trazem tantas alegrias todos os dias. E por permitir que eu faça parte de sua família.

Ao Cleiton, sempre tão presente, por todo o amor, pela compreensão, por perdoar a minha ausência durante esta pesquisa, mas principalmente por todo o companheirismo e por toda a diversão, te agradeço por estar presente em todos os melhores momentos da minha vida.

Aos meus amigos, Ednei, Franciane, Hannah, Leonardo e Rodrigo, por terem feito parte da minha graduação, do meu mestrado, e a partir daí, de toda a minha vida. Obrigada pelos conselhos, almoços, festas, conversas, sempre tão bem vindas. Aos meus companheiros de mestrado, Camila, Danilo, Jorge, Mariana e Tiago, pela amizade construída, por transformarem os longos períodos de estudos, os extensos trabalhos e o dia a dia de pesquisas, muito mais proveitosos e divertidos.

Ao meu coorientador Alvim, pela sabedoria, direção e por todos os ensinamentos tão importantes e indispensáveis para o desenvolvimento desta pesquisa. À professora Deise, pela orientação, incentivo e apoio, e por sempre me receber com tanto bom humor fazendo com que eu confiasse cada vez mais em meu trabalho.

A todos os professores e colegas do PPGMNE pelos ensinamentos. Ao Instituto Tecnológico SIMEPAR pela oportunidade de desenvolvimento desta pesquisa. A todos que contribuíram para a realização deste trabalho: o meu muito obrigada.

RESUMO

Este trabalho tem como objetivo principal detectar possíveis anomalias em séries de dados de vazão. A análise da qualidade de dados hidrológicos é extremamente importante, pois todos os dados observados (monitorados) necessitam de tratamentos e processamentos básicos para que possam ser utilizados com confiabilidade. Por meio de técnicas utilizadas para a resolução de problemas de previsão e classificação, baseadas em redes neurais *Self-Organizing Maps* (SOM), *Radial Basis Functions* (RBF), e métodos de interpolações (*smooth spline*) de dados, busca-se apontar possíveis anomalias nas séries oriundas dos postos hidrológicos das sub-bacias de Porto Amazonas e União da Vitória do estado do Paraná, fornecidos pelo Sistema Meteorológico do Paraná (SIMEPAR). Os três métodos propostos retornaram resultados satisfatórios, cumprindo o objetivo da pesquisa, entretanto, o projeto utilizado para a aplicação da rede neural RBF demonstrou capacidade superior de detecção de anomalias nas séries de vazão, em especial para a sub-bacia de Porto Amazonas que é considerada uma sub-bacia de resposta rápida a ocorrência de precipitação.

Palavras-chave: Vazão. Anomalias. Redes Neurais. Interpolação.

ABSTRACT

This work aims to detect possible anomalies in data flow series. Quality analysis of hydrological data is extremely important, because all the observed (monitored) data require basic treatments and processing so they can be used reliably. Through techniques used for solving prediction and classification problems, based on neural networks Self-Organizing Maps (SOM), Radial Basis Functions (RBF), and methods of interpolation (smooth spline) data, it was seeked to identify possible anomalies in the series from the hydrological stations of the sub-basins of Porto Amazonas and União da Vitória in Paraná state, provided by the Sistema Meteorológico do Paraná (SIMEPAR). The three proposed methods returned satisfactory results, fulfilling the purpose of the research, however, the project used for the application of RBF neural network demonstrated superior ability to detect anomalies in flow series, especially for the sub-basin of Porto Amazonas which is considered a rapid response sub-basin to precipitation events.

Key-words: Flow. Anomalies. Neural Networks. Interpolation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ANA	Agência Nacional de Águas
AUC	Area Under the Curve
CQ	Controle de Qualidade
DDA	Dynamic Decay Adjustment
EE	Exponencial Estendido
GML	Generalized Maximum Likelihood
LVQ	Learning Vector Quantization
MLP	Multi Layer Perceptron
RBF	Radial Basis Functions
RCE	Restricted Coulomb Energy
ROC	Relative Operating Characteristic
SIMEPAR	Sistema Meteorológico do Paraná
SOM	Self-Organizing Maps
UFPR	Universidade Federal do Paraná

SUMÁRIO

1 INTRODUÇÃO

A observação de informações hidrológicas, tal como o nível da água (cota) em rios, pode ser realizada em postos de monitoramento automáticos ou convencionais, em que são registrados os valores observados e armazenados em um banco de dados. Além dos dados coletados, existem os que são calculados, por exemplo, a vazão. Através deste banco são construídas longas séries extremamente importantes para a confecção de estudos hidrológicos, que também podem ser utilizados em operações de empreendimentos, que envolvem recursos hídricos, dos mais diversos tipos entre os quais os empreendimentos energéticos.

Entretanto, neste monitoramento podem ocorrer falhas causadas por erros provocados pelo equipamento de medição, falhas na transmissão dos dados, depredação dos equipamentos, entre outros, implicando na inclusão de dados inconsistentes às séries, ou ainda, causando a ausência de informações em alguns momentos. Assim, possíveis incertezas na medição das cotas podem resultar em valores anômalos para as variáveis de vazão estimadas.

Para suprimir estas falhas, o Sistema Meteorológico do Paraná (SIMEPAR) desenvolveu métodos específicos de análise de consistência para identificação de dados espúrios através de um processo normalmente denominado de Controle de Qualidade (CQ), porém, apenas erros grosseiros são identificados ao passo que registros inconsistentes menos discrepantes são negligenciados. Portanto, posteriormente ao CQ, as séries de dados são inspecionadas pelo corpo técnico através da análise gráfica de curtos intervalos ao longo de todo o período de dados, no intuito de identificar inconsistências não reconhecidas pelo CQ. Contudo, a análise gráfica exige do profissional tempo, atenção e conhecimento suficientes para que todas as anomalias sejam reconhecidas.

Para auxiliar no processo de consistência dos dados, este trabalho abordará a utilização de técnicas para a "detecção de anomalias" voltadas à previsão e classificação de dados. Para a previsão de dados em uma série, utiliza-se a técnica de mapas auto-organizáveis (*Self-Organizing Maps*- SOM) das redes neurais de Kohonen e o método de interpolação de dados *Splines* e sua variante denominada neste trabalho como *Smooth Spline*. Para a classificação dos dados que compõe uma série de vazão em anômalos ou normais, utilizam-se as redes neurais de funções de base radial (*Radial Basis Functions*- RBF) e o algoritmo de ajuste de decaimento dinâmico (*Dynamic Decay Adjustment*- DDA) para redes RBF.

Estas técnicas serão aplicadas com o intuito de reconhecer padrões de comportamento nas séries de dados, identificando períodos com comportamentos singulares, e assim, apontar a um profissional capacitado em consistência de séries possíveis anomalias nos dados.

A fim de constatar a eficiência destas técnicas, seus resultados serão confrontados entre si, e com as séries de dados verificadas por técnicos experientes em consistência do SIMEPAR. Para este fim apresentam-se os métodos de avaliação da qualidade dos modelos probabilísticos que envolvem matrizes de contingência e curvas de características operacionais relativas (*Relative Operating Characteristic*- ROC) e por fim, aplicou-se o Teorema de Bayes, visando a geração das probabilidades posteriores condicionadas a um indicador.

Como referência aos dados de vazão, adotou-se os postos hidrológicos pertencentes à bacia do rio Iguaçu das cidades de União da Vitória e Porto Amazonas do estado do Paraná.

1.1 JUSTIFICATIVA

A obtenção de séries de dados hidrológicos altamente confiáveis, acrescida de valiosas informações sobre seu comportamento, pode ajudar operadores e planejadores do sistema energético e de diversos outros sistemas em suas funções diárias,

proporcionando precisão às suas decisões e, por consequência, eficiência em suas operações. Além disso, estes dados possuem grande importância para a sociedade, pois por meio deles são feitos monitoramentos de eventos considerados críticos, como por exemplo, cheias e estiagens, entre outros.

Pretende-se facilitar o trabalho dos profissionais que realizam a consistência das séries tornando-as úteis aos sistemas de previsão de vazão, através dos métodos matemáticos estudados e da construção de um sistema de apoio ou suporte à decisão na identificação de dados anômalos em séries de vazão.

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

Este trabalho tem como objetivo geral desenvolver, através da utilização de Redes Neurais Artificiais ou métodos de interpolação de dados, uma metodologia para apontar dados anômalos em séries de vazão dos postos hidrológicos de União da Vitória e Porto Amazonas no estado do Paraná para que se possa utilizá-la, futuramente, nas demais estações hidrológicas monitoradas pelo SIMEPAR. Pretende-se com esta metodologia facilitar o trabalho de consistência de séries de dados de vazão.

1.2.2 OBJETIVOS ESPECÍFICOS

- Construir um projeto de aplicação para o SOM a fim de identificar anomalias através da previsão de uma série de dados de vazão;
- Construir um projeto de aplicação para o *Smooth Spline* a fim de identificar anomalias através da previsão de uma série de dados de vazão;
- Construir um projeto de aplicação para a RBF com a utilização do algoritmo DDA a fim de identificar anomalias através da classificação de uma série de dados de vazão;

- Promover a comparação entre os projetos de aplicação através de validação dos métodos a fim de selecionar o mais eficiente;
- Utilizar o projeto selecionado como facilitador no trabalho de consistência realizado pelo SIMEPAR.

1.3 LIMITAÇÕES DO TRABALHO

Esta dissertação limita-se a propor metodologias que apontem dados anômalos em séries de vazão, de forma que não serão abordados métodos para a correção, tão pouco para previsão destes dados, ainda que os métodos SOM e *Smooth Spline* proponham uma sugestão dos valores a serem substituídos nas séries. Posteriormente pretende-se utilizar estes métodos em um sistema de suporte à decisão que será elaborado em trabalhos futuros.

1.4 ESTRUTURA DO TRABALHO

Este trabalho está dividido em 6 capítulos, incluindo-se esta introdução.

O Capítulo 2 destina-se à descrição do problema proposto, dos dados utilizados neste trabalho, dos tipos de anomalias que podem ser detectadas nestes dados, bem como da localização da região de estudo.

No Capítulo 3 são apresentadas algumas referências de trabalhos sobre os métodos a serem utilizados: *Self-Organizing Maps* (SOM), *Smooth Spline* e *Radial Basis Functions* (RBF), seguidas de suas fundamentações teóricas e da descrição do algoritmo de decaimento dinâmico (*Dynamic Decay Adjustment* - DDA), das curvas ROC (Relative Operating Characteristic), e do Teorema de Bayes.

No Capítulo 4 são descritos os projetos de aplicação construídos para cada um dos métodos utilizados.

No Capítulo 5 propõe-se um método de avaliação da acurácia e validação dos

modelos propostos para a previsão e classificação dos dados. Além disso, são apresentados os resultados das aplicações.

O capítulo 6 destina-se às considerações finais do trabalho, bem como algumas sugestões para trabalhos futuros .

2 DESCRIÇÃO DO PROBLEMA

Este capítulo destina-se à descrição do problema a ser estudado neste trabalho, discute-se o tipo e a importância dos dados hidrológicos e como eles são coletados, os diferentes tipos de inconsistências encontrados nas séries de dados de cota, a influência destas inconsistências sobre as anomalias dos dados de vazão, e a localização da região de estudo.

2.1 DADOS HIDROLÓGICOS

Existem diversos tipos de dados hidrológicos, entre eles, os dados sobre cotas, vazões, chuvas, evaporação, perfil do rio, qualidade da água e sedimentos, entre outros (ANA, 2013). Segundo a Agência Nacional de Águas (ANA), a coleta destes dados é de extrema importância para a sociedade, pois é utilizada para produzir estudos, definir políticas públicas e avaliar a disponibilidade hídrica. Por meio dessas informações, a ANA, o SIMEPAR, e demais órgãos, monitoram eventos considerados críticos, como cheias e estiagens, disponibilizam informações para a execução de projetos, identificam o potencial energético, de navegação ou de lazer em um determinado ponto ou ao longo do manancial, levantam as condições dos corpos de água para atender a projetos de irrigação ou de abastecimento público, entre outros.

Devido a essa grande importância, nesta pesquisa, trabalha-se com séries de dados de vazão de rios que são calculados através dos dados de cota (valor do nível do rio observado através de réguas de medição) coletados em postos hidrológicos. Em qualquer processo de modelagem, a familiaridade com os dados disponíveis é indispensável e o pré-processamento dos dados tem efeito significativo no desempenho do modelo. Apesar dos modelos baseados em redes neurais serem considerados capazes de trabalhar com dados ruidosos e incompletos, segundo Valença (2005)

o pré-processamento dos dados pode melhorar substancialmente o desempenho de qualquer modelo.

A coleta dos dados de cota de um rio ou reservatório é realizada em postos de monitoramento convencionais ou automáticos (FIGURA 1). Convencionalmente, a série é coletada por um leiturista normalmente às 7h e às 17h, através de observações em réguas, registradas e enviadas para os institutos para a formação de um banco de dados. Na coleta automática, os valores de cota do rio, em metros, são coletados através de transdutores de pressão em pequenos intervalos de tempo. O SIMEPAR armazena em seu banco de dados, valores de cota a cada 15 minutos.

(A) (B)

FIGURA 1: Postos de monitoramento: (A) Convencional e (B) Automático
FONTE: A autora (2014)

Através da série de dados de cota, é calculada a série de vazão, que representa o volume de um fluido que atravessa uma dada seção do escoamento por unidade de tempo, e é estimada em m^3/s. Nas seções monitoradas pelos postos hidrológicos do SIMEPAR a vazão dos rios é estimada a partir da medição de cota no próprio posto, e esta estimativa é realizada com o uso de curvas de descarga que estabelecem uma relação unívoca entre cota e vazão, portanto assim são calculados os dados de vazão

dos diversos postos hidrológicos em particular, de Porto Amazonas - PR (primeiro objeto deste estudo). Porém, em União da Vitória - PR (segundo objeto deste estudo), devido a efeitos de remanso, as vazões são obtidas por algoritmos mais complexos, por ser necessário o registro de cota em Porto Vitória - PR, além da cota no próprio posto, para obter a vazão do rio Iguaçu.

Assim, a vazão é calculada, principalmente, para a operação dos reservatórios, porém, especialmente neste processo, as incertezas da variável cota, que podem ocorrer devido às falhas nos equipamentos de medição, falhas na transmissão dos dados, depredação dos equipamentos, entre outros, podem resultar em valores inconsistentes (neste trabalho denominados anômalos) para a variável vazão estimada.

2.2 CONSISTÊNCIA

Como citado no Capítulo 1, o SIMEPAR aplica métodos específicos de análise de consistência para identificação de dados espúrios com inconsistências grosseiras através do CQ, e posteriormente as séries de dados são inspecionadas pelo corpo técnico do SIMEPAR, no intuito de identificar inconsistências não reconhecidas por métodos anteriores. A Figura 2 exemplifica os mais comuns tipos de anomalias em dados de vazão, apresentados a seguir através de um levantamento feito por Breda e Negrão (2012):

- Mudanças de *Offsets*: geralmente os sensores que realizam o monitoramento da cota da água nos postos hidrológicos são do tipo célula de pressão. Entretanto, com o passar do tempo, alguns fatores ambientais podem alterar a resposta do equipamento. Quando há uma diferença entre os registros do sensor e os dados da leitura de régua é aplicado um fator compensatório aos dados do sensor antes de eles serem armazenados no banco de dados. Este fator compensatório é a simples adição, ou subtração, da diferença de cota entre a leitura manual e a automática, e é denominado de *offset*. Porém, a aplicação deste *offset*, que pode ser para cima (aumentando os valores dos dados de cota) ou para baixo

(diminuindo os valores dos dados de cota), nem sempre traz benefícios à consistência dos dados, pois eles podem ser alterados indevidamente, sendo necessário apontar e posteriormente desfazer estas modificações. Neste trabalho procura-se utilizar os métodos para apontar apenas os dados extremos de uma mudança de *offset*, os quais são considerados dados anômalos (os dados centrais às mudanças de *offset* são considerados normais), para efeito de análise dos métodos apresentados.

- *Spikes*: são erros em que um registro, ou um curto período de registros, fica deslocado, para cima ou para baixo, da tendência da série, formando picos ou vales muito sinuosos, repentinos, que fogem do padrão de comportamento da série. Pretende-se apontar os dados relativos a estes *spikes* como anômalos através dos métodos.

- Oscilações Diárias: ocorrem devido à oscilação na carga da bateria que alimenta o sensor de pressão utilizado para registrar o nível da água (a cota) nos rios. Notou-se que oscilações intensas ocorrem mais frequentemente em períodos de recessão, onde a cota do rio está baixa, ou em períodos noturnos no caso de sensores alimentados por energia solar. Resultam em fortes oscilações nos valores dos dados que fogem ao comportamento normal da série, pretende-se apontar, com os métodos aplicados, alguns dados pertencentes a estes períodos de oscilações significativas.

- Ruídos: eventualmente as séries de cotas analisadas apresentam ruídos em certos períodos. Não foi buscada uma explicação precisa sobre o porquê da ocorrência deste tipo de falha, porém sabe-se que elas ocorrem, pois são oscilações evidentes fugindo do comportamento normal das séries, apontam-se como anomalias alguns valores destes períodos da mesma maneira que para as oscilações diárias.

- Falhas: ausência de dados que podem acontecer devido à falta de medição em um determinado período de tempo, seja pela ausência do leiturista no posto

hidrológico, pela falta de bateria nos sensores, inutilização dos sensores por diversas razões, entre outros motivos.

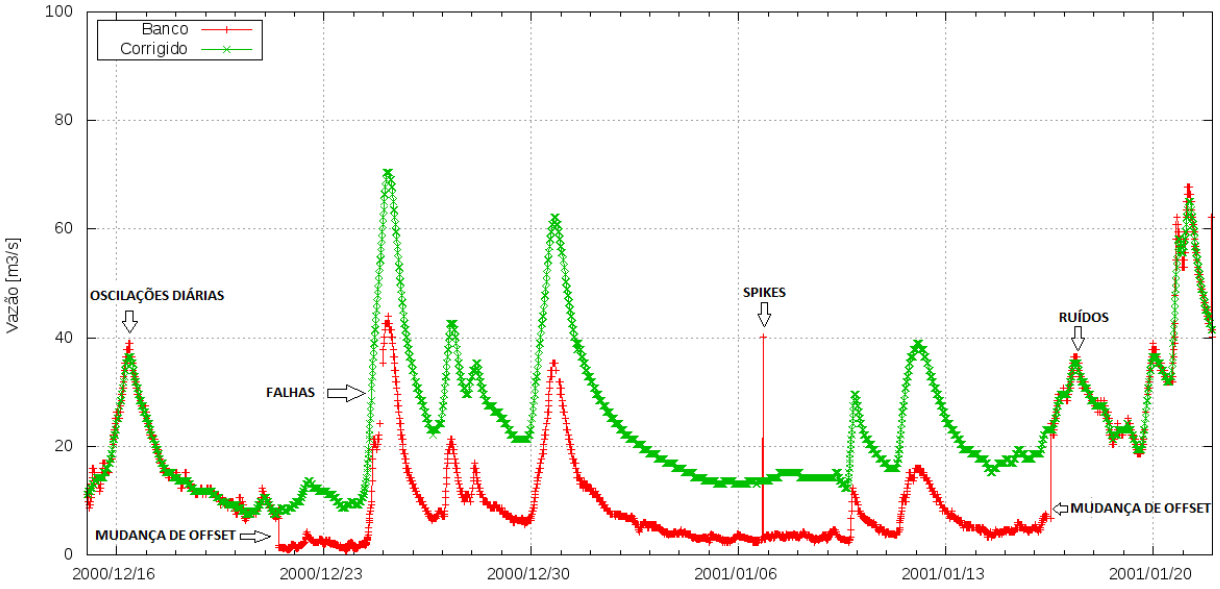


FIGURA 2: Exemplos de inconsistências

FONTE: A autora (2014)

Estes diversos tipos de inconsistências ocorridas em séries de dados de cotas implicam diretamente em ocorrências de anomalias nas séries de dados de vazões. Portanto, a menos dos problemas de falhas que já estão automaticamente identificados nos bancos de dados compostos por séries de vazão, o objetivo da aplicação dos métodos propostos por esta pesquisa (SOM, *Smooth Spline* e RBF-DDA) às séries de dados de União da Vitória e Porto Amazonas é alertar os valores limites de uma mudança de *offset* mal sucedida, os *spikes*, para baixo ou para cima, isolados ou em conjunto, e os momentos em que oscilações diárias e ruídos ocorreram na série de dados.

2.3 LOCALIZAÇÃO

Os dados de vazão a serem analisados são provenientes dos postos hidrológicos das sub-bacias de União da Vitória e Porto Amazonas, ambos localizados na bacia do

rio Iguaçu no estado do Paraná. “Para a consistência de dados de uma bacia hidrográfica é essencial o reconhecimento da sua área e a sua dinâmica hídrica” (NEGRÃO, 2011).

O mapa apresentado na Figura 3 mostra a bacia do rio Iguaçu e seus postos de monitoramento hidrológicos. Pode-se notar que a sub-bacia de União da Vitória é interna à bacia do rio Iguaçu estando sobre a influência das demais sub-bacias à montante, e a sub-bacia de Porto amazonas é de cabeceira, não sofrendo influência de outras sub-bacias.

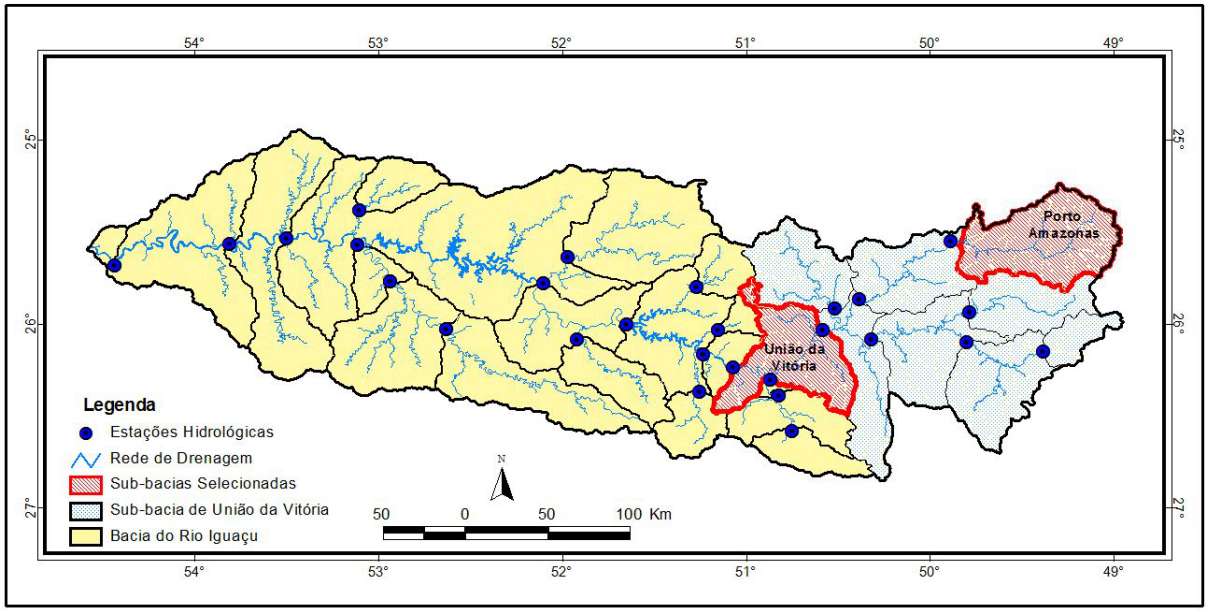


FIGURA 3: Bacia do rio Iguaçu e seus postos hidrológicos
FONTE: Adaptada de Breda (2013)

A sub-bacia de União da Vitória está localizada em um ponto central da bacia do rio Iguaçu, sendo assim, pode sofrer influência das sub-bacias que estão localizadas à montante (lado contrário à corrente de água) bem como influenciar as sub-bacias que estão localizadas à jusante (lado para onde se dirige a corrente de água), além disso, a região da sub-bacia de União da Vitória, onde localizam-se as sedes municipais de União da Vitória, Porto União e Porto Vitória, abrange as cidades mais severamente afetadas por inundações no estado, causadas diretamente pela vazão do leito principal do rio Iguaçu, que compreende uma sub-bacia de aproximadamente 25.000 km^2 (JICA,

1995a *apud* LEITE, 2008).

Leite (2008) enfatiza que “as inundações na região de União da Vitória são extensivas, severas, envolvem prejuízos significativos para a população e economia locais, são condicionadas pelas regras operativas da usina hidrelétrica de Foz do Areia, localizada à jusante das cidades afetadas, e devem requerer medidas estruturais e não estruturais para sua mitigação.” Assim considera-se que esta região merece atenção, principalmente com relação aos dados de vazão.

A sub-bacia de Porto Amazonas está localizada no início da bacia do rio Iguaçu, ou seja, é uma sub-bacia de cabeceira e, portanto apenas influencia o comportamento das sub-bacias que estão localizadas à jusante. A região da sub-bacia de Porto Amazonas corresponde ao município de Curitiba, região metropolitana e entornos, e possui aproximadamente 4.000 km^2 (NEGRÃO, 2011).

Por pertencer a uma região de área mista (parte urbanizada e parte rural) a sub-bacia de Porto Amazonas possui comportamento bastante irregular, e é conhecida como uma sub-bacia de resposta rápida das vazões com relação à ocorrência de precipitação. Além disso, a alta densidade demográfica, o uso irregular do território urbano e o desmatamento da vegetação ciliar são alguns dos fatores preponderantes para a contaminação da bacia, cuja disponibilidade hídrica tem sido colocada à prova, além de acarretar em problemas frequentes de alagamentos nas épocas de chuva forte.

Analisando a curva de autocorrelação das duas sub-bacias citadas (FIGURA 4), observam-se valores altos para a sub-bacia de União da Vitória, indicando como alta a sua regularidade, no sentido de que não existem variações significativas em seus dados de cota, e como consequência em seus dados de vazão ao longo de um grande período de tempo, assim ela pode ser considerada como uma sub-bacia de resposta lenta das vazões com relação à ocorrência de chuva. Em contrapartida, os menores valores de autocorrelação da sub-bacia de Porto Amazonas demonstra que ela pode ser considerada menos regular em relação à União da Vitória, característica das sub-

bacias de cabeceira.

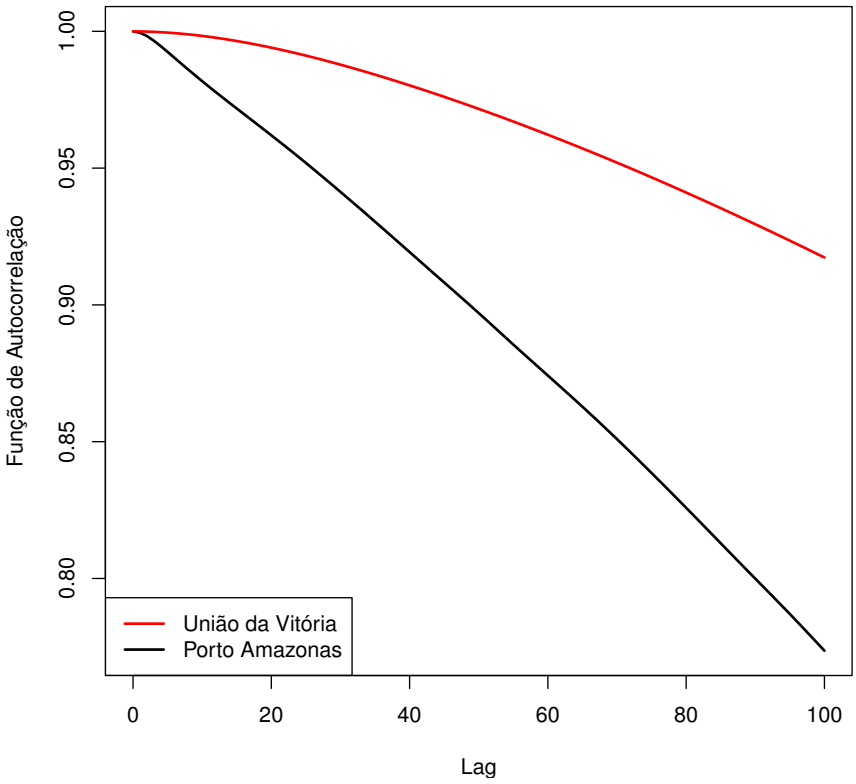


FIGURA 4: Comparação das autocorrelações das sub-bacias de União da Vitória e Porto Amazonas

FONTE: A autora (2014)

Sendo assim, estes são os principais motivos das escolhas destas localidades para esta pesquisa, já que pretende-se aplicar os métodos estudados às sub-bacias de comportamentos bastante distintos a fim de avaliar se os modelos são capazes de detectar anomalias em diferentes situações.

3 TÉCNICAS DE PREVISÃO E CLASSIFICAÇÃO APLICADAS EM SÉRIES DE DADOS PARA DETECÇÃO DE ANOMALIAS

Este capítulo apresenta um levantamento de referências bibliográficas, bem como a descrição de métodos utilizados neste trabalho para a detecção de anomalias organizados em dois grupos: (1) previsão de dados em uma série, com a utilização da técnica de mapas auto-organizáveis (*Self-Organizing Maps* - SOM) das redes neurais de Kohonen e do método de interpolação de dados *Splines* e sua variante denominada *Smooth Spline*; (2) classificação em uma série de dados, com a utilização das redes neurais de funções de base radial (*Radial Basis Functions* - RBF) e do algoritmo de ajuste de decaimento dinâmico (*Dynamic Decay Adjustment* - DDA) para redes RBF. Além disso, apresentam-se os métodos de avaliação da qualidade dos modelos probabilísticos que envolvem matrizes de contingência e curvas ROC (*Relative Operating Characteristic*) e por fim, a aplicação do Teorema de Bayes, utilizado para a geração das probabilidades posteriores condicionadas a um indicador.

3.1 TÉCNICAS BASEADAS EM PREVISÃO DE DADOS

3.1.1 SELF-ORGANIZING MAPS

Redes neurais artificiais, ou simplesmente redes neurais são modelos que tomam por base as redes neurais biológicas associadas ao processamento paralelo do cérebro humano que possui bilhões de neurônios capazes de se interconectar, de processar milhões de informações e realizar milhões de ligações sinápticas. De maneira geral uma rede neural artificial é um sistema constituído por elementos de processamento interconectados, chamados de neurônios, que estão dispostos em camadas, de entrada, intermediárias e de saída, e são responsáveis pela não-linearidade e pela memória adquirida pela rede (VALENÇA, 2005).

Existem diversos tipos de aplicações de redes neurais nas mais variadas áreas de pesquisa, assim como, existem diversos tipos de redes neurais. Neste trabalho, especificamente, utiliza-se uma das redes propostas por Kohonen (2001) conhecida como *Self-Organizing Maps* (SOM), ou mapas auto-organizáveis. A seguir serão apresentados os mais diversos tipos de aplicações desta rede encontrados em alguns trabalhos de pesquisa, e as definições e configurações deste tipo de rede. Além disso, no Capítulo 4 será apresentada uma metodologia para utilizar esta rede neural.

Devido ao avanço dos estudos envolvendo redes neurais, uma técnica de aprendizado competitivo e não supervisionado foi desenvolvida com uma nova formalização na distribuição das camadas de neurônios (MONTGOMERY; JR, 2007), Kohonen (2001) a denominou de *Self-Organizing Maps* (SOM), neste mesmo livro o autor propôs ainda variações para esta técnica, como o LVQ (*Learning Vector Quantization*) similar ao SOM, porém descrevendo uma aprendizagem supervisionada, e sem a ordenação espacial dos neurônios da camada de saída realizada pelo SOM. Entre outras variações.

Desde que Teuvo Kohonen propôs esta técnica, estudos dos mais diversos tipos, nas mais diversas áreas foram realizados. Os principais objetivos destas pesquisas se concentram em reconhecer padrões em séries de dados, problemas de classificação, problemas de visualização e otimização.

Em seu artigo, Vesanto e Alhoniemi (2000) relatam a importância do SOM como ferramenta na fase exploratória de mineração de dados (*data-mining*), apresentando diferentes abordagens para agrupamento a partir do SOM. Em particular, o uso de agrupamento aglomerativo hierárquico e o agrupamento partitivo utilizando k-médias são investigados. O processo é proposto em duas fases, na primeira utiliza-se o SOM para produzir protótipos que são então clusterizados em um segundo estágio, assim obtém-se um melhor desempenho em relação a uma clusterização mais direta, sem o tratamento inicial dos dados, além de reduzir o tempo computacional.

Em uma nova abordagem, Siqueira (2005) utilizou o SOM a fim de determinar,

em uma fase inicial do problema de designação linear, a matriz de custos, e para a resolução do problema propriamente dito utilizou-se de uma segunda rede neural denominada Rede Neural Recorrente de Wang com a aplicação do princípio de *Winner Takes All*. O autor enfatiza que a fase de definição da matriz de custos é de extrema importância, pois sem ela a solução final do problema não seria a ideal. Aplicou-se esta metodologia em um estudo de caso: o problema de alocação de salas de aula para disciplinas de graduação e pós-graduação da Universidade Federal do Paraná (UFPR), onde mapas com diversas dimensões para a determinação dos custos deste problema foram testados. Os resultados encontrados foram considerados satisfatórios, com baixa taxa de erro.

A técnica SOM é utilizada para a classificação e reconhecimento de padrões no artigo de Cavazos (1999), a fim de obter uma ferramenta de diagnóstico e reconhecimento de padrões atmosféricos anômalos característicos de eventos extremos de precipitação e auferir a precipitação diária na região estudada durante o período de 1980 a 1993. Assim, aplicou-se o SOM com o intuito de encontrar características importantes que caracterizassem os campos de umidade e circulação atmosférica diária (por exemplo, os padrões dominantes atmosféricos) durante o inverno sobre a área em estudo. Esta pesquisa demonstrou que as anomalias climáticas na região, típicas das condições extremas, e classificadas pelo SOM, são fisicamente compatíveis com os padrões locais que têm sido utilizados. Além disso, a técnica SOM oferece a vantagem de revelar várias fontes de variação em diferentes escalas de tempo.

3.1.1.1 DESCRIÇÃO DO MÉTODO

A técnica de agrupamento e visualização *Self-Organizing Maps* (SOM) é um tipo de rede neural artificial de *Kohonen* treinada através de aprendizagem não supervisionada, para produzir uma classificação própria dos dados de entrada (preservando a sua estrutura) que possuem características comuns entre si.

Segundo Castro (2006) na estrutura desta rede os neurônios estão dispostos em

nós de uma grade que normalmente é uni ou bidimensional, como ilustrado na Figura 5. No caso de um mapa bidimensional, a geometria é livre, podendo ser quadrada, retangular, triangular, hexagonal, etc. representando o tipo de vizinhança dos neurônios (FIGURA 6).

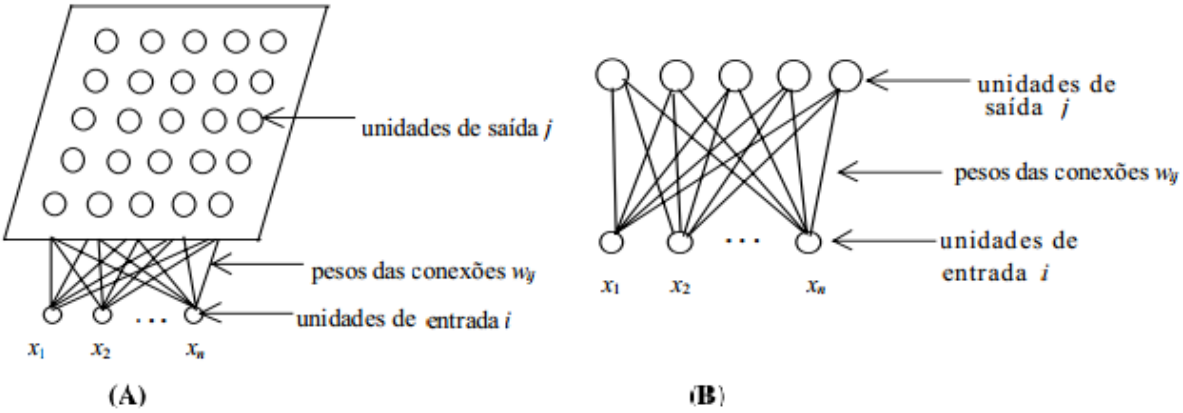


FIGURA 5: Arquitetura do SOM: (A) Grade da saída bidimensional e (B) Grade da saída unidimensional

FONTE: Adaptada de Castro (2006)

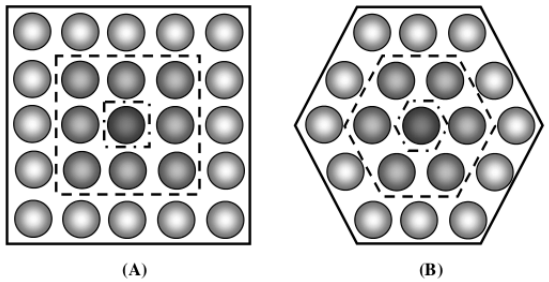


FIGURA 6: Tipos de vizinhança: (A) Quadrada e (B) Hexagonal

FONTE: Siqueira (2005)

Esta rede é treinada através de um esquema de aprendizagem competitiva, ou seja, ao se apresentar uma entrada à rede os neurônios competem entre si e o vencedor tem seus pesos ajustados (MONTGOMERY; JR, 2007), portanto, existem conexões laterais entre as unidades de saída que não são mostrados na Figura 5.

Existe ainda, um processo de cooperação entre o neurônio vencedor e seus vizinhos topológicos (definidos por um raio de vizinhança), pois ambos recebem ajustes, sendo assim, as características estatísticas intrínsecas de um dado de entrada in-

fluenciam na arquitetura da rede, ou seja, as unidades de saída do SOM tornam-se topologicamente ordenadas e assim, os neurônios vizinhos correspondem a regiões similares no espaço de entrada (KOHONEN, 2001).

3.1.1.2 O ALGORITMO DA REDE

Segundo Montgomery e Jr (2007) o processo se inicia através da escolha arbitrária de pequenos valores aleatórios dos pesos sinápticos $(w_{ij}) = [w_{i1}, w_{i2}, \ldots, w_{in}]^T$, a fim de que não seja imposta ao mapa uma organização prévia. Em seguida realizam-se os processos de competição, cooperação e adaptação sináptica.

• Processo Competitivo:

1. Apresenta-se um vetor de entrada $x = [x_1, x_2, \ldots, x_n]^T$ selecionado aleatoriamente dentre as demais unidades de treinamento;

2. Calcula-se a distância entre o vetor de entrada e o vetor de pesos dos neurônios: $d_i(t) = \sum_{j=1}^{N}(x_j(t) - w_{ij}(t))^2$; em que $d_i(t)$ é a distância euclidiana quadrática entre o vetor de pesos do neurônio i e o vetor de entrada na iteração t; i é o índice do neurônio, j é o índice do nó de entrada; N é a quantidade de dados de entrada, ou seja, a dimensão do vetor x; $x_j(t)$ é a entrada no nó j da iteração t; $w_{ij}(t)$ é o valor do peso sináptico entre o nó de entrada j e o neurônio i na iteração t;

3. Seleciona-se o neurônio vencedor através da menor distância resultante.

• Processo Cooperativo:

O centro de uma vizinhança topológica de neurônios cooperativos é indicado pelo neurônio vencedor, assim, o parâmetro vizinhança topológica h_{ik}, que indica o grau de cooperação entre o neurônio i e seu vizinho k, é simétrico em relação ao neurônio vencedor k e decresce monotonamente com o aumento da distância lateral l_{ik} até que, no limite em que l_{ik} tende a infinito e h_{ik} tende a zero. A fim de respeitar estas

condições utiliza-se a função gaussiana (EQUAÇÃO 1):

$$h_{ik} = e^{(-\frac{l_{ik}^2}{2\sigma^2})} \tag{1}$$

na qual o termo l_{ik}^2 representa a distância euclidiana entre os neurônios i e k, o parâmetro σ representa a largura efetiva da vizinhança topológica que diminui com o passar do tempo, implicando assim em uma diminuição dos valores de h_{ik} caracterizando uma vizinhança mais restrita e pode ser representado pela Equação 2:

$$\sigma(t) = \sigma_0 \cdot e^{-\frac{t}{\tau_1}} \tag{2}$$

em que σ_0 é o valor inicial de σ, t é o número de iterações e τ_1 é uma constante de tempo.

• Processo Adaptativo:

O ajuste de um peso sináptico w_{ij}, entre o nó de entrada j e o neurônio i, é representado pela Equação 3:

$$\Delta w_{ij} = \alpha(t) \cdot h_{ik}(t) \cdot (x_j - w_{ij}) \tag{3}$$

onde o parâmetro vizinhança topológica na iteração t é $h_{ik}(t)$ e $\alpha(t)$ é a taxa de aprendizagem, com valores entre zero e um, e geralmente tem sido definida como:

$$\alpha(t) = \alpha_0 \cdot e^{-\frac{t}{\tau_1}} \tag{4}$$

na qual α_0 é o valor inicial adotado. O principal objetivo desta taxa, decrescente ao longo do tempo, é evitar o supertreinamento ocasionado pela apresentação excessiva de dados novos de entrada após um longo treinamento.

3.1.2 INTERPOLAÇÃO - *SPLINE* E *SMOOTH SPLINE*

Interpolar uma função $f(x)$ consiste em aproximá-la por uma outra função $g(x)$, escolhida entre uma classe de funções definida a priori e que satisfaça algumas propriedades (RUGGIERO; LOPES, 1996). A função $g(x)$ é então utilizada em substituição à

função $f(x)$ quando, por exemplo, somente alguns valores numéricos são conhecidos e deseja-se calcular o valor da função em um ponto não tabelado, ou quando a função possui uma expressão complicada dificultando os cálculos de diferenciação e integração, ou ainda como no caso deste trabalho em que os valores numéricos da função são conhecidos, porém deseja-se verificar a sua validade, efetuando uma espécie de "previsão" de um dado já existente.

Existem diversas formas de se interpolar polinomialmente, através da fórmula de Taylor, por polinômios de Hermite, pelas formas de Lagrange, de Newton, resolvendo sistemas lineares, entre outros. Porém, segundo Ruggiero e Lopes (1996) nenhum desses métodos possui convergência garantida, e esta é a maior vantagem de se utilizar funções *spline* em problemas de interpolação, pois desta maneira, interpola-se $f(x)$ em grupos de poucos pontos, obtendo-se polinômios de menores graus, e impondo-se condições para que a função de aproximação seja contínua e tenha derivadas contínuas até uma determinada ordem.

Definição 3.1.1 *Considera-se a função $f(x)$ tabelada nos pontos $x_0 < x_1 < \cdots < x_n$. Uma função $S_p(x)$ é denominada* spline *de grau p com nós nos pontos x_i, $i = 0, 1, \ldots, n$, se satisfizer as seguintes condições:*

1. *em cada subintervalo $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, (n-1)$, $S_p(x)$ é um polinômio de grau p: $s_p(x)$.*

2. *$S_p(x)$ é contínua e tem derivada contínua até ordem $(p-1)$ em $[x_0, x_n]$.*

   Se, além disto, $S_p(x)$ também satisfaz a condição:

3. *$S_p(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$, então será denominada* spline *interpolante.*

3.1.2.1 *SPLINE* LINEAR INTERPOLANTE

A função *spline* linear interpolante de $f(x)$, $S_1(x)$, nos nós $x_0, x_1, \ldots, x_n$ pode ser escrita em cada subintervalo $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, n$ como

$$s_i(x) = f(x_{i-1})\frac{x_i - x}{x_i - x_{i-1}} + f(x_i)\frac{x - x_{i-1}}{x_i - x_{i-1}}, \qquad \forall x \in [x_{i-1}, x_i]. \tag{5}$$

Pois:

1. $S_1(x)$ é um polinômio de grau 1 em cada subintervalo $[x_{i-1}, x_i]$, por definição;
2. $S_1(x)$ é contínua em (x_{i-1}, x_i), por definição, e, nos nós x_i, S_1 está bem definida, pois: $s_i(x_i) = s_{i+1}(x_i) = f(x_i) \Rightarrow S_1(x)$ é contínua em $[x_0, x_n]$ e, portanto, $S_1(x)$ é *spline* linear;
3. $S_1(x_i) = s_i(x_i) = f(x_i) \Rightarrow S_1(x)$ é *spline* linear interpolante de $f(x)$ nos nós $x_0, \ldots, x_n$.

Uma desvantagem na utilização da *spline* linear é que sua derivada primeira é descontínua nos nós. Se *splines* quadráticas forem utilizadas, $S_2(x)$ possui derivadas contínuas até ordem 1 apenas, e portanto, a curvatura de $S_2(x)$ pode ser invertida nos nós. Sendo assim, utilizam-se *splines* cúbicas com mais frequência.

3.1.2.2 *SPLINE* CÚBICA INTERPOLANTE

Uma *spline* cúbica, $S_3(x)$, é uma função polinomial por partes contínua onde cada parte, $s_k(x)$, é um polinômio de grau 3 no intervalo $[x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$. A curva de $S_3(x)$ não possui picos e não troca de curvatura nos nós pois sua primeira e segunda derivadas são contínuas.

Supondo que $f(x)$ esteja tabelada nos pontos x_i, $i = 0, 1, 2, \ldots, n$ a função $S_3(x)$ é chamada *spline* cúbica interpolante de $f(x)$ nos nós x_i, $i = 0, 1, 2, \ldots, n$ se existem n polinômios de grau 3, $s_k(x)$, $k = 1, 2, \ldots, n$ tais que:

1. $S_3(x) = s_k(x)$ para $x \in [x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$

2. $S_3(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$

3. $s_k(x_k) = s_{k+1}(x_k)$, $k = 1, 2, \ldots, (n-1)$

4. $s'_k(x_k) = s'_{k+1}(x_k)$, $k = 1, 2, \ldots, (n-1)$

5. $s''_k(x_k) = s''_{k+1}(x_k)$, $k = 1, 2, \ldots, (n-1)$

E por simplicidade $s_k(x) = a_k(x - x_k)^3 + b_k(x - x_k)^2 + c_k(x - x_k) + d_k,\ k = 1, 2, \ldots, n$, portanto o cálculo de $S_3(x)$ exige a determinação de 4 coeficientes para cada k, em um total de $4n$ coeficientes: a_1, b_1, c_1, d_1, a_2, b_2, $\ldots$, a_n, b_n, c_n, d_n.

3.1.2.3 *SMOOTH SPLINE*

Existe ainda, a possibilidade de se interpolar sequências de dados através de diferentes tipos de *splines*, ou seja, por meio de uma técnica denominada *Smooth Spline* em que diferentes escolhas de um parâmetro de suavidade da função implicam em interpolações mais ou menos suaves dos dados, de acordo com a necessidade de resolução de cada problema.

Huang *et al.* (2013) utilizaram *Smooth Spline* juntamente com correção residual para fundir observações pluviométricas e dados de sensoriamento remoto, a fim de construir campos de precipitação em grades diárias com altas resoluções espaciais. Em primeiro lugar, as medidas observadas foram utilizadas como variável resposta. *Smooth Spline* e um campo de precipitação em grades existentes foram utilizados como variáveis explicativas para estimar a tendência da superfície de precipitação. Uma abordagem para estimar a matriz de covariância de erro do campo de precipitação interpolado também foi fornecida. E, por fim, um conjunto de dados de precipitação diários observados da Nova Zelândia foram aplicados para validar a abordagem proposta. Os resultados sugeriram que a abordagem de interpolação proposta produziu superfícies de precipitação com alta resolução espacial e menores erros de interpolação em ambos os dados esparsos e dados de áreas densas.

Smooth Spline é uma técnica usual para ajuste de curva, em que a seleção do parâmetro de suavidade é crucial. Este foi o fato que motivou Chen e Huang (2010) a estudar diferentes métodos como: Mallows'C_p, *generalized maximum likelihood* (GML), e o critério exponencial estendido (EE), para selecionar o melhor parâmetro. Através da comparação entre eles, os autores propuseram uma adaptação para C_p que permite a construção de uma expressão analítica que pode ser facilmente calculada. Demonstrou-se que esta adaptação se tornou superior e mais estável que os demais métodos.

O trabalho de Kouibiaa e Pasadas (2008) aborda o problema da construção de algumas curvas de forma livre, e superfícies, a partir de diferentes tipos de dados: exatos e ruidosos. Os autores estenderam a teoria dos $D^m - Splines$ ao longo de um domínio limitado para dados ruidosos até os vetores variacionais de *Smooth Spline*, e obtiveram resultados de convergência para ambos os dados (ruidosos e exatos), além disso apresentaram algumas estimativas dos erros.

Segundo Fox (2002) *Smooth Splines* são utilizados a fim de solucionar o seguinte problema de regressão simples: encontrar uma função $\hat{f}(x)$ com primeira e segunda derivadas contínuas que minimize a penalização da soma dos quadrados,

$$SS^*(h) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + h \int_{x_{min}}^{x_{max}} [f''(x)]^2 dx \tag{6}$$

em que h é um parâmetro de suavização do *spline*.

O primeiro termo da Equação 6 representa a soma dos quadrados dos resíduos. O segundo termo é uma *penalidade de rugosidade*, que é grande quando a segunda derivada integrada da função de regressão $f''(x)$ é grande, isto é, quando $f(x)$ muda de declive rapidamente, além disso, a integração é sobre o intervalo de x (os dados a serem interpolados).

Em um extremo, quando a constante de suavização é definida como $h = 0$ (e se todos os valores de x são distintos), $\hat{f}(x)$ simplesmente interpola os dados, porém a curva resultante pode se tornar muito ruidosa já que pretende-se passar por todos os

pontos. Em outro extremo, se h é muito grande, então $\hat{f}$ é selecionada de modo que $\hat{f}''(x)$ é sempre nula, o que equivale a uma regressão local com vizinhança infinita, porém esta pode ser uma aproximação suave demais, o que implica em uma curva que não satisfaz o formato dos dados fugindo da realidade.

A função $\hat{f}(x)$ que minimiza a Equação 6 é um *spline* cúbico natural (que inclui dois nós adicionais nas extremidades dos dados, e restringe a função para que ela seja linear além destes pontos) com nós nos diferentes valores observados de x. Embora este resultado pareça indicar que n parâmetros são necessárias (quando todos os valores de x são distintos), a "penalidade de rugosidade" impõe restrições adicionais à solução, reduzindo substancialmente o número equivalente de parâmetros para o *Smooth Spline* prevenindo $\hat{f}(x)$ da interpolação dos dados. De fato, é comum a escolha do parâmetro de suavização h indiretamente, definindo o número equivalente de parâmetros do *Smooth Spline* de modo que ele se torne o mais suave possível.

Selecionando o Parâmetro de Suavização:

O método *Smooth Spline* possui um parâmetro de suavização ajustável. Este parâmetro pode ser selecionado pelo julgamento visual e através de cálculos de erros, escolhendo um valor que equilibre a suavidade e a fidelidade aos dados. Métodos mais formais de seleção de parâmetros de suavidade normalmente tentam minimizar o erro médio quadrático do ajuste, quer pelo emprego de uma fórmula de aproximação, ou por alguma forma de validação cruzada. Na validação cruzada, os dados são divididos em subgrupos (possivelmente compreendendo as observações individuais), o modelo é ajustado sucessivamente omitindo cada subconjunto por sua vez, e em seguida, o modelo ajustado é utilizado para "prever" a resposta para o próximo subconjunto. Através do teste de diferentes valores de parâmetros de suavidade encontrar-se-á um valor que minimiza a validação cruzada da estimativa do erro médio quadrático, porém, devido ao fato de validação cruzada exigir extremo esforço computacional, aproximações são frequentemente utilizadas (FOX, 2002).

3.2 TÉCNICAS BASEADAS EM CLASSIFICAÇÃO DE DADOS

3.2.1 REDE NEURAL DE FUNÇÃO DE BASE RADIAL (*RADIAL BASIS FUNCTION* - RBF)

As redes neurais de função de base radial (*Radial-Basis Function* - RBF) são capazes de resolver problemas do tipo classificação, aproximação de funções e previsão de séries temporais. O rápido treinamento de uma RBF é a grande vantagem para a resolução destes tipos de problemas (OLIVEIRA, 2004).

Em seu trabalho, Kindelan e Bayona (2013) exploraram a aplicabilidade do método RBF para modelar a propagação da chama laminar. Este problema foi considerado um desafio interessante para o método RBF, pois envolveu a solução de duas equações lineares parabólicas acopladas de temperatura e fracção de massa. Mostrou-se a adequação do método de resolução de problemas de propagação de chama instáveis uni e bidimensionalmente, e também aplicou-se o método para calcular a forma de uma chama ancorada usando ambos os nós equidistantes e não equidistantes.

Redes neurais do tipo RBF foram utilizadas juntamente com resultados de exames eletrocardiogramas por Xianhai (2011) para que se pudesse estudar o reconhecimento de emoções. Compararam-se os métodos de reconhecimento de padrão emocional entre as redes neurais classificadoras do tipo *Backpropagation* e RBF e se efetuou a comparação de seus resultados experimentais. A classificação das amostras através da rede neural *Backpropagation* recebeu taxas de reconhecimento global de 87,5%, já para a RBF a taxa de reconhecimento geral foi de 91,67%. Assim, em comparação com a rede neural do tipo *Backpropagation*, a RBF apresentou uma melhor taxa de desempenho para o reconhecimento de padrões emocionais.

Há algum tempo pesquisas tem demonstrado que as técnicas de redes neurais podem ser utilizadas com sucesso para a estimativa de chuvas a partir das medições de radar (LIU; CHANDRASEKAR; XU, 2001). A RBF pode ser utilizada como um método não paramétrico para representar a relação entre as medições de radar e a taxa de precipitação obtida através de medições pluviométricas. A eficácia da estimativa

de chuva, usando redes neurais pode ser influenciada por vários fatores, como a representatividade e suficiência do conjunto de dados de treinamento, a capacidade de generalização da rede para novos dados, mudança de estação, mudança de local, e assim por diante. Liu, Chandrasekar e Xu (2001) apresentaram um novo esquema de forma adaptativa atualizando a estrutura e os parâmetros da rede neural para a estimativa de chuvas. Os dados coletados por um Radar-1988 Doppler (WSR- 88D) e uma rede de pluviômetros foram utilizados para avaliar o desempenho da rede adaptativa para a estimativa de chuvas. Mostra-se que a rede adaptativa pode estimar chuvas com bastante precisão.

Segundo o manual Zell *et al.* (1998) o princípio da função de base radial deriva da teoria da aproximação funcional, em que dados N pares $(\vec{x}_i, y_i)$ $(x \in R^n$ e $Y \in R)$, procura-se uma função f da forma:

$$f(\vec{x}) = \sum_{i=1}^{K} c_i h(|\vec{x} - \vec{t}_i|) \tag{7}$$

onde h é a função de base radial, $\vec{t}_i$ são os K centros da função de base radial que devem ser previamente definidos, os coeficientes c_i são os pesos desconhecidos a principio e devem ser calculados, $\vec{x}_i$ e $\vec{t}_i$ são elementos de um vetor espacial n dimensional, e aplica-se a função h à distância euclidiana entre cada centro $\vec{t}_i$ e argumento de entrada $\vec{x}$. Geralmente a função h é representada por uma função gaussiana e para este caso, valores de $\vec{x}$ que são próximos do centro $\vec{t}$ produzem um valor de saída da rede igual a um, enquanto que longas distâncias entre a entrada $\vec{x}$ e o centro $\vec{t}$ retornam uma saída aproximadamente nula. A função f deve ser uma aproximação para os N pares dados $(\vec{x}_i, y_i)$ e deve minimizar a seguinte função de erro H:

$$H[f] = \sum_{i=1}^{N} (y_i - f(\vec{x}_i))^2 + \lambda ||Pf||^2 \tag{8}$$

A primeira parte da definição de H (a soma) é a condição que minimiza o erro total da aproximação, isto é, que obriga f a aproximar os N pontos dados. A segunda parte de H $(||Pf||^2)$ é um estabilizador que força f a se tornar tão suave quanto seja

possível. O fator λ determina a influência do estabilizador.

Sob certas condições é possível mostrar que um conjunto de coeficientes c_i podem ser calculados de modo que H se torne mínimo. Este cálculo depende dos centros $\vec{t}_i$ que devem ser escolhidos antecipadamente. Se os seguintes vetores e matrizes forem introduzidos: $\vec{c} = (c_1, c_2, \ldots, c_K)^T$, $\vec{y} = (y_1, y_2, \ldots, y_N)^T$,

$$G = \begin{pmatrix} h(|\vec{x}_1 - \vec{t}_1|) & \ldots & h(|\vec{x}_1 - \vec{t}_K|) \\ \vdots & \ddots & \vdots \\ h(|\vec{x}_N - \vec{t}_1|) & \ldots & h(|\vec{x}_N - \vec{t}_K|) \end{pmatrix} \text{ e } G_t = \begin{pmatrix} h(|\vec{t}_1 - \vec{t}_1|) & \ldots & h(|\vec{t}_1 - \vec{t}_K|) \\ \vdots & \ddots & \vdots \\ h(|\vec{t}_K - \vec{t}_1|) & \ldots & h(|\vec{t}_K - \vec{t}_K|) \end{pmatrix}$$

então o conjunto de parâmetros desconhecidos c_i pode ser calculado pela fórmula:

$$\vec{c} = (G^T \cdot G + \lambda G_t)^{-1} \cdot G^T \cdot \vec{y} \tag{9}$$

O método das funções de base radial pode ser representado por uma rede neural de três camadas. O funcionamento deste tipo de rede difere significativamente de outras redes como, por exemplo, o SOM citado anteriormente, pois contém uma camada intermediária com funções de bases radiais e a ativação dos neurônios desta camada se dá através do cálculo da distância entre o vetor de entrada e um vetor protótipo, no sentido de que quando um padrão de entrada é apresentado à rede, a distância entre o neurônio central e este padrão é calculada, e a saída da rede para o neurônio central é o resultado da aplicação da função de base radial a esta distância. A Figura 7 mostra a arquitetura de uma rede RBF.

Nesta arquitetura:

- A camada de entrada possui n unidades (os elementos do vetor $\vec{x}$);
- Os K componentes da soma na definição de f são representados pelas unidades da camada intermediária;
- $\vec{t}_i$ representam as ligações entre os padrões de entrada e as unidades ocultas;
- As unidades da camada intermediária são representadas pelo cálculo da distân-

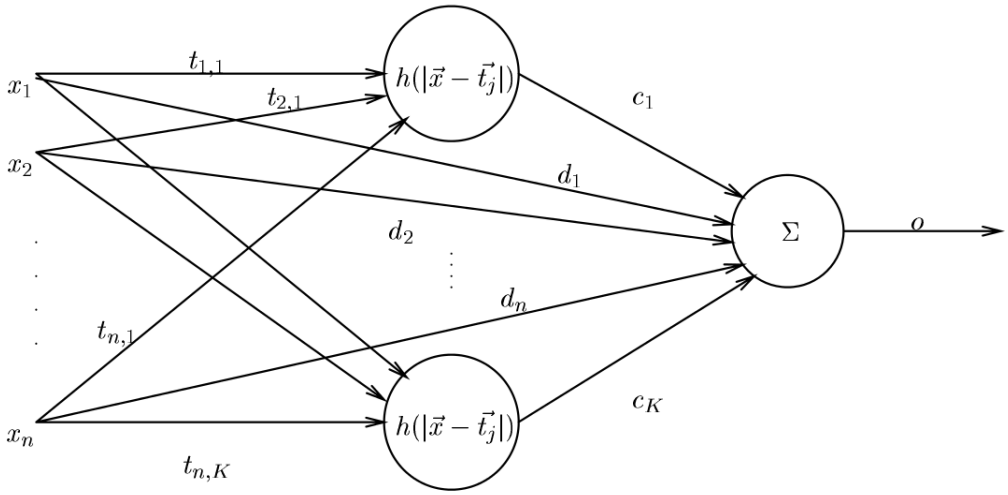


FIGURA 7: Arquitetura da RBF
FONTE: Zell *et al.* (1998)

cia euclidiana entre os padrões de entrada e as ligações $\vec{t_i}$ equivalentes à cada entrada;

- A ativação das unidades da camada intermediária é dada pela aplicação da função h à distância euclidiana calculada;
- O único neurônio de saída recebe contribuição de todos os neurônios da camada oculta;
- As ligações que conectam a camada intermediária ao neurônio de saída são representadas pelos coeficientes c_i;
- A ativação do neurônio de saída é determinada pela soma ponderada de suas chegadas.

Esta arquitetura pode ser expandida com a modificação de algumas características: aumentar o número de neurônios de saída, ativar as unidades de saída através de uma função σ (sigmoide) não-linear inversível, adicionar um *bias* relacionado ao neurônio de saída, conexões diretas entre entrada e saída, utilizar um *bias* nas unidades da camada intermediária. Com estas características, uma rede neural totalmente

conectada é capaz de representar o seguinte conjunto de aproximações:

$$o_k(\vec{x}) = \sigma\left(\sum_{j=1}^{K} c_{j,k} h\left(|\vec{x}-\vec{t}_j|, p_j\right) + \sum_{i=1}^{n} d_{i,k} x_i + b_k\right) = \sigma\left(f_k\left(\vec{x}\right)\right),\ k = 1,\ldots,m \tag{10}$$

Esta fórmula descreve o comportamento de uma rede do tipo *feedforward* completamente conectada, com n entradas, K neurônios ocultos e m saídas. $o_k(\vec{x})$ é a ativação do neurônio de saída k para a entrada $(\vec{x})$. Os coeficientes $c_{j,k}$ representam as ligações entre a camada intermediária e a de saída. $d_{i,k}$ são as ligações diretas entre entrada e saída. Os *bias* das unidades de saída são os b_k, e os da camada intermediária são os p_j que determinam as características da função h. Por fim, a função de ativação dos neurônios de saída é representada por σ.

Zell *et al.* (1998) enfatiza que a grande vantagem do método de funções de base radial é a possibilidade do um cálculo direto dos coeficientes $c_{j,k}$ (as ligações entre a camada intermediária e de saída) e do *bias* b_k. Este cálculo requer uma escolha adequada dos centros $\vec{t}_j$ (as ligações entre a entrada e a camada intermediária). Devido à falta de conhecimento sobre a qualidade de $\vec{t}_j$, é recomendado realizar alguns ciclos de treinamento da rede depois do cálculo direto dos pesos. Uma vez que os pesos das ligações entre a entrada e a camada de saída também não podem ser calculados diretamente, deve haver um processo de treinamento especial para redes neurais que utilizam as funções de base radial.

O procedimento de treinamento implementado para este trabalho tenta minimizar o erro E usando gradiente descendente. Recomenda-se a utilização de diferentes taxas de aprendizagem para diferentes grupos de parâmetros de treinamento. O seguinte conjunto de fórmulas contém todas as informações necessárias para o treinamento:

$$E = \sum_{k=1}^{m}\left(\sum_{i=1}^{N}\left(y_{i,k} - o_k\left(\vec{x}_i\right)\right)^2\right), \quad \Delta\vec{t}_j = -\eta_1 \frac{\partial E}{\partial \vec{t}_j}, \quad \Delta p_j = -\eta_2 \frac{\partial E}{\partial p_j},$$

$$\Delta c_{j,k} = -\eta_3 \frac{\partial E}{\partial c_{j,k}}, \quad \Delta d_{i,k} = -\eta_3 \frac{\partial E}{\partial d_{i,k}}, \quad \Delta b_k = -\eta_3 \frac{\partial E}{\partial b_k}$$

Uma melhoria usual para o procedimento de treinamento é a definição de um erro

máximo permitido no interior dos neurônios de saída. Isso impede que a rede fique supertreinada, já que os erros que são menores do que os valores predefinidos são tratados como zero impedindo que as ligações correspondentes sejam alteradas.

A cargo de simplificação, Oliveira (2004) ressalta que redes RBF são redes neurais *feedforward* (todas as conexões têm a mesma direção, partindo da camada de entrada rumo à camada de saída) com uma única camada intermediária onde não existem pesos associados a esta camada e a camada de entrada. As unidades da camada oculta utilizam funções de base radial (RBF) cujo valor aumenta ou diminui em relação à distância de um ponto central, e as funções mais utilizadas nestas redes são as funções gaussianas, em que cada unidade da camada intermediária calcula uma saída R_i através de:

$$R_i(\vec{x}) = exp(-\frac{||\vec{x}-\vec{r}_i||^2}{\sigma_i^2}) \tag{11}$$

onde $\vec{x}$ é o vetor de entrada, $||\vec{x}-\vec{r}||$ é a distância euclidiana entre o vetor de entrada $\vec{x}$ e o centro da função gaussiana $\vec{r}_i$, e σ_i é o raio da função de base radial. Os neurônios da camada oculta são ligados aos da camada de saída por ligações ponderadas. A última camada calcula a saída para cada classe da seguinte forma:

$$f(\vec{x}) = \sum_{i=1}^{m} A_i \cdot R_i(\vec{x}) \tag{12}$$

em que m é o número de neurônios na camada intermediária e A_i são os pesos das conexões i.

O teorema de Cover sobre a separabilidade de padrões encontrado em Haykin (1998) justifica o uso das RBFs, pois afirma que um problema de classificação padrão não-linear em um espaço de alta dimensionalidade tem maior probabilidade de ser linearmente separável neste espaço do que em um outro de baixa dimensionalidade, e geralmente uma rede RBF possui um grande número de unidades na camada oculta justamente com a intenção de aumentar a dimensão do espaço em que se está trabalhando, a fim de tornar o problema linearmente separável.

Um grande número de técnicas de treinamento têm sido utilizadas, e Oliveira

(2004) cita algumas delas. Baseando-se nele, este trabalho utiliza redes do tipo RBF que serão treinadas utilizando o algoritmo de Ajuste de Decaimento Dinâmico (*Dynamic decay adjustment* - DDA), um algoritmo de treinamento construtivo juntamente com alguns dos métodos de aperfeiçoamento deste algoritmo propostos por Oliveira (2004).

3.2.2 AJUSTE DE DECAIMENTO DINÂMICO (*DYNAMIC DECAY ADJUSTMENT* - RBF-DDA)

O algoritmo de ajuste de decaimento dinâmico (DDA) é uma extensão do algoritmo RCE (*Restricted Coulomb Energy*) (HUDAK, 1992 *apud* OLIVEIRA, 2004) e propicia um treinamento fácil, construtivo e rápido para redes neurais de função de base radial (ZELL *et al.*, 1998). Resultados experimentais apresentados na literatura mostram que o DDA supera o RCE e outros algoritmos de treinamento para redes RBF em um grande número de problemas de classificação (OLIVEIRA, 2004).

A arquitetura de uma RBF treinada com o algoritmo DDA (RBF-DDA) pode ser observada na Figura 8. O número de unidades na camada de entrada representa a dimensionalidade do espaço de entrada, esta camada é totalmente ligada à camada intermediária que é única e seu número de unidades é determinado automaticamente durante o treinamento, estas unidades ocultas são representadas por funções de ativação gaussianas. A saída da RBF-DDA é a classificação resultante da rede representada por classes de 1 até n. O problema de classificação utiliza a abordagem vencedor leva tudo, segundo a qual a unidade com o maior ativação determina a classe. Cada unidade da camada intermediária é ligada a uma única unidade de saída e estas ligações possuem pesos A_i, as unidades de saída são formadas por funções de ativação lineares com valores calculados através da Equação 12.

Na Figura 8 o vetor de pesos que liga as unidades de entrada a todas as unidades intermediárias representa o centro da função gaussiana. A distância euclidiana entre o vetor de entrada e o vetor da camada oculta (ou protótipo) é usada como uma entrada para a função gaussiana que resulta em uma resposta local, se o vetor de entrada for

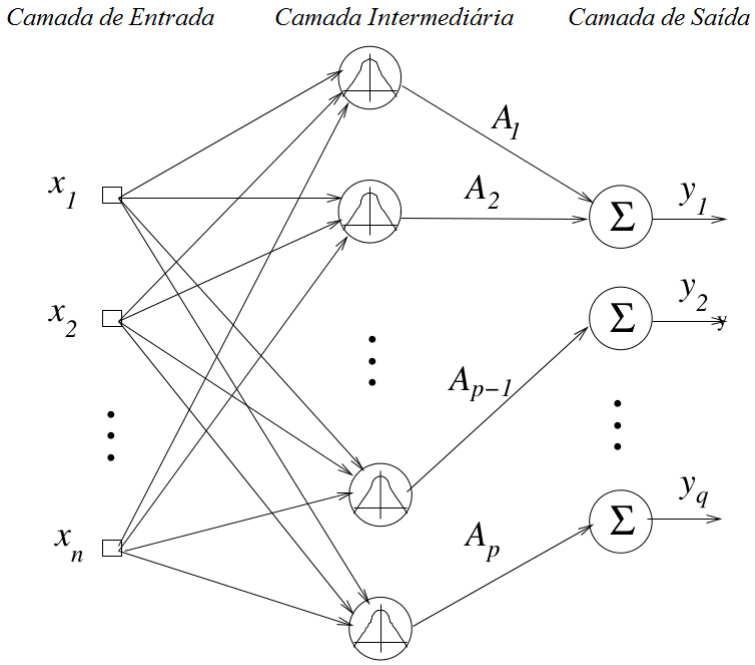


FIGURA 8: A estrutura de uma RBF-DDA.
FONTE: Oliveira (2004)

próximo ao protótipo, a unidade terá uma ativação elevada, caso contrário a ativação será próxima de zero. Cada unidade de saída calcula uma soma ponderada de todas as ativações das unidades ocultas pertencentes à classe correspondente.

O algoritmo de treinamento DDA é construtivo, pois inicia com a camada intermediária sem neurônios e eles são adicionados quando necessário no decorrer do treinamento, além disso, os centros das unidades da camada intermediária $\vec{r}_i$, e seus raios σ_i, são determinados pelo DDA durante o treinamento. Os valores dos pesos das ligações entre a camada escondida e de saída também são obtidos através do algoritmo.

A decisão da introdução de novos neurônios à camada intermediária da rede depende de dois parâmetros. Um deles é o limite positivo θ^+, que deve ser ultrapassado pela ativação de um neurônio da mesma classe para que nenhum novo protótipo seja adicionado. O outro é um limite negativo θ^-, que é o limite superior para a ativação de classes conflitantes, ou seja, o limite superior para que um neurônio de uma nova classe seja adicionado (ZELL *et al.*, 1998). O uso de dois limites resulta em uma

melhor classificação em áreas onde o algoritmo não introduz novos neurônios. Um exemplo está representado na Figura 9, nele, um novo padrão de entrada da classe B, classificado corretamente, resulta em ativações acima do limite positivo para a classe correta B e abaixo do limite negativo para a classe conflitante A (OLIVEIRA, 2004).

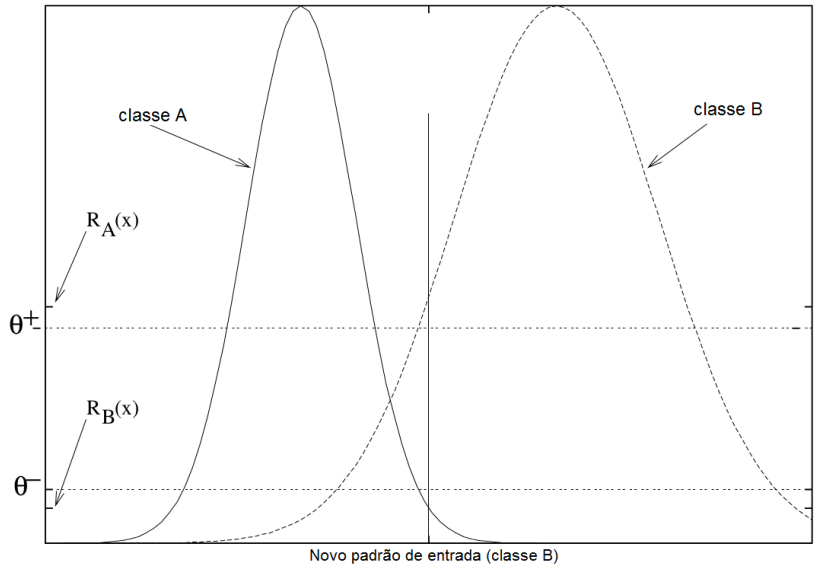


FIGURA 9: Classificação de um novo padrão da classe B por RBF-DDA.
FONTE: Adaptada de Oliveira (2004)

Uma rede RBF-DDA treinada satisfaz as seguintes equações para todo padrão de entrada $\vec{x}$ de uma classe c (correta):

$$\exists i : R_i^c(\vec{x}) \geq \theta^+ \tag{13}$$

$$\forall k \neq c, 1 \leq j \leq m_k : R_j^k(\vec{x}) \leq \theta^- \tag{14}$$

Nota-se que as condições acima não garantem a classificação correta de todos os padrões de treinamento, pois são satisfeitas para unidades da camada intermediária e não para as unidades de saída.

O algoritmo DDA é executado até que não sejam detectadas alterações nos valores dos parâmetros (número de unidades na camada oculta e seus respectivos parâmetros e valores de pesos). Isso geralmente ocorre em apenas 4 ou 5 iterações (ZELL *et al.*, 1998). Este critério de parada natural evita o supertreinamento dos dados. Em

cada iteração o algoritmo começa definindo todos os pesos como zero, pois, caso contrário, eles iriam acumular informações duplicadas sobre os padrões de treinamento.

O algoritmo cria um novo protótipo para um determinado padrão de treinamento $\vec{x}$ somente se não houver algum protótipo da mesma classe na rede cuja saída seja $R_i(\vec{x}) \geq \theta^+$. Caso contrário, o algoritmo apenas aumenta o peso A_i da ligação associada com uma das unidades ocultas (da mesma classe do padrão de treinamento) que resulta $R_i(\vec{x}) \geq \theta^+$.

Quando um novo protótipo é introduzido à rede, o seu centro tem o mesmo valor do vetor de treinamento $\vec{x}$ e o peso da sua ligação com a camada de saída é definido com valor 1. O raio da função gaussiana é escolhido de tal modo que os resultados produzidos pelo novo protótipo para protótipos existentes das classes conflitantes são menores do que θ^-. E então, há uma fase de retração, na qual os raios dos protótipos em conflito são ajustados para produzir valores de saída menores que θ^- para o padrão de treinamento $\vec{x}$.

A Figura 10 apresenta um pequeno exemplo do algoritmo DDA em ação. O primeiro padrão a partir do conjunto de treino é da classe A, o algoritmo DDA cria um protótipo centrado sobre este padrão, como mostrado na Figura 10 (1). O seguinte padrão é da classe B, como mostrado na Figura 10 (2), isto implica na introdução de um novo protótipo de classe B e também na redução do protótipo da classe anterior A. Na Figura 10 (3) o terceiro padrão de treinamento, também da classe B, é apresentado, o DDA não introduz um novo protótipo vez que já existe um protótipo cuja saída para esse novo padrão é maior do que θ^+, no entanto, é necessário diminuir o protótipo da classe A, pois caso contrário, iria produzir um valor maior do que θ^- para um padrão da classe B. Finalmente, na Figura 10 (4) um novo protótipo para a classe A é introduzido, pois a saída do protótipo existente desta classe é menor do que θ^+ para o quarto padrão de treinamento.

Resultados experimentais utilizando diferentes tipos de dados têm demonstrado que o desempenho de generalização da RBF-DDA depende ligeiramente dos valores

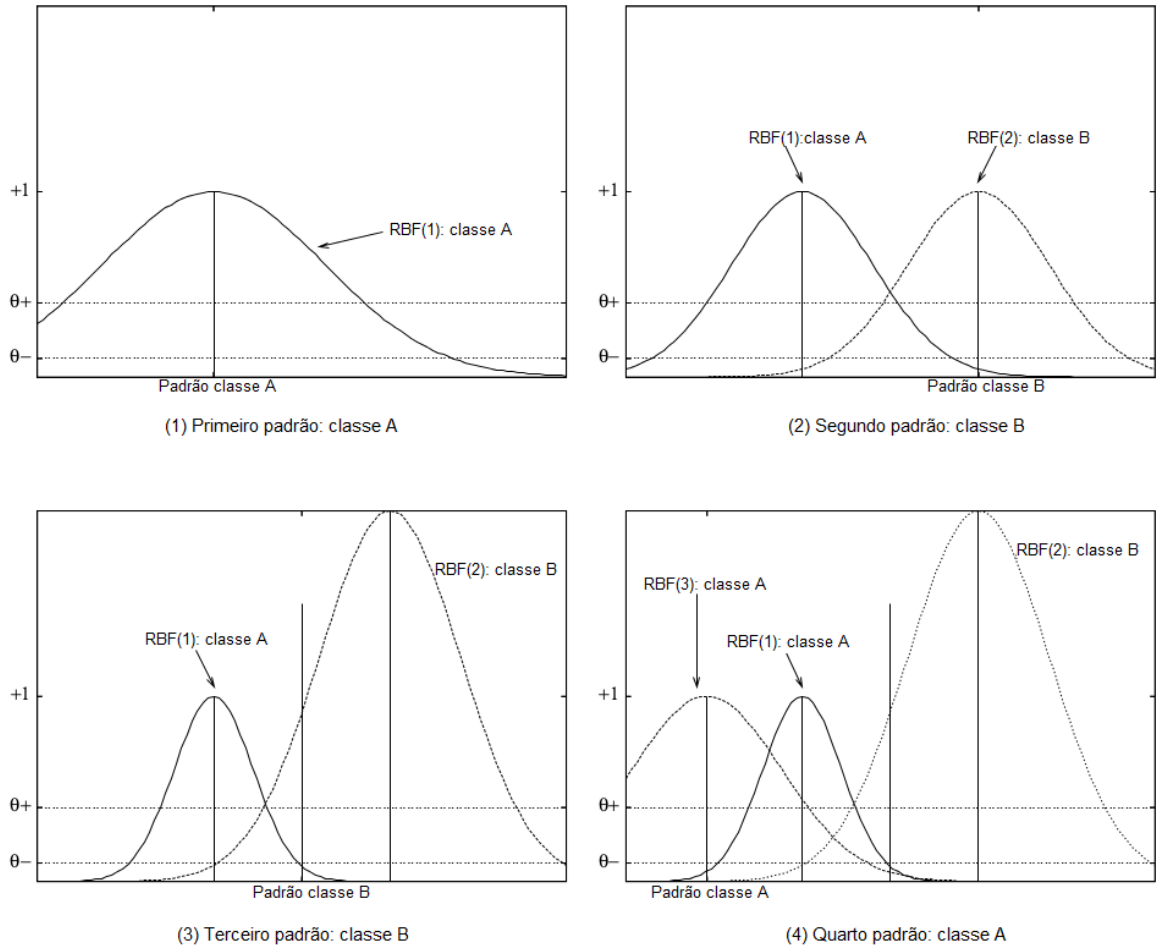


FIGURA 10: Exemplo do algoritmo DDA
FONTE: Adaptada de Oliveira (2004)

dos parâmetros θ^+ e θ^- (OLIVEIRA, 2004). Isto levou Berthold e Diamond, criadores do algoritmo DDA para RBF, a acreditar que esses parâmetros não eram críticos para o desempenho de problemas de classificação. Sendo assim, recomenda-se a utilização de valores padrões para esses parâmetros, $\theta^+ = 0.4$ e $\theta^- = 0.2$ (BERTHOLD; DIAMOND, 1995 *apud* OLIVEIRA, 2004) para qualquer conjunto de dados. No entanto Oliveira (2004) observou que para alguns conjuntos de dados importantes, tais como dados de reconhecimento de imagem, menores valores de θ^- resultam em uma melhoria considerável no desempenho, o que lhe motivou a propor métodos para aprimorar o desempenho da RBF-DDA, entre eles a variação do parâmetro θ^-, a técnica de treinamento negativo, e a separação do conjunto de treinamento a fim de se utilizar um conjunto de validação. Sua tese (OLIVEIRA, 2004) apresenta diversos métodos e

este trabalho se utiliza de alguns deles para o problema em estudo.

3.2.2.1 APRIMORANDO A RBF-DDA ATRAVÉS DE TREINAMENTO NEGATIVO

A técnica de treinamento negativo consiste em treinar a rede neural com um conjunto de treinamento ampliado formado por padrões de treinamento originais juntamente com um conjunto de amostras negativas (OLIVEIRA, 2004) formado por dados anômalos de diferentes tipos, e situados em diferentes contextos. Pretende-se com a utilização desta técnica aprimorar o uso da RBF-DDA a partir do seu treinamento, pois a ideia do método é criar um atrator no espaço de entrada, de modo que quando um novo padrão aleatório (uma anomalia) é apresentado à rede após o treinamento espera-se classificá-lo como uma anomalia e não como um membro válido de uma classe de treinamento.

O problema é que, em aplicações práticas, é muito difícil selecionar as amostras negativas desejáveis, então é comum a utilização de um grande número de amostras negativas aleatórias. Algumas experiências têm indicado que isso, algumas vezes, retorna bons resultados na detecção de anomalias, porém em outros casos, essa melhora não ocorre.

O método proposto por Oliveira (2004) treina redes do tipo MLP (*Multi Layer Perceptron*) e RBF, além disso, o autor ressalta que qualquer tipo de modelo classificador pode ser utilizado em conjunto com este método, atribuindo amostras negativas para o vetor de saída relacionado aos dados anômalos. Nesta pesquisa esta técnica será apresentada no Capítulo 4 a fim de utilizá-la no treinamento da RBF-DDA para o problema em estudo. A Figura 11 apresenta este método.

A rede neural classificadora é treinada com um conjunto de treinamento ampliado contendo o mesmo número de padrões anômalos e padrões considerados normais. Para que um padrão aleatório seja considerado normal ele não deve se desviar muito do padrão original pelo qual foi gerado.

Para treinar e testar a rede neural procede-se da seguinte forma:

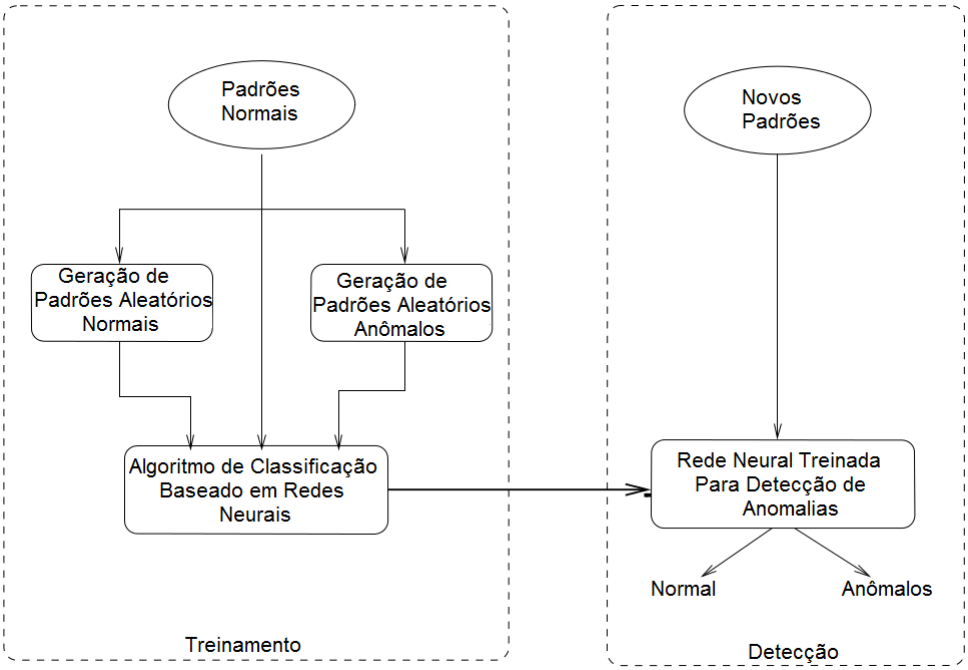


FIGURA 11: Método de detecção de anomalias com amostras negativas
FONTE: Adaptada de Oliveira (2004)

- Para cada padrão normal disponível geram-se n padrões normais aleatórios;
- Para cada padrão normal disponível geram-se $n+1$ padrões anômalos aleatórios;
- Cria-se um conjunto ampliado de padrões incluindo os padrões normais originais, os padrões normais e anômalos gerados aleatoriamente. Divide-se o conjunto de padrões ampliados resultantes em dois novos conjuntos disjuntos de treinamento e de teste;
- Treina-se uma rede neural para classificação de padrões utilizando o conjunto de treinamento ampliado;
- Avalia-se o desempenho do sistema com o uso do conjunto de teste ampliado.

Deve-se notar que um conjunto de teste ampliado também é gerado a fim de avaliar o desempenho do método, pois, geralmente a série de tempo utilizada para o período de teste, nas experiências relatadas, não apresenta um número significativo de

dados anômalos, sendo assim têm-se as anomalias que se deseja simular, gerando o conjunto de teste ampliado da mesma maneira que os conjuntos de treinamento ampliados foram gerados.

3.2.2.2 GERAÇÃO DE CONJUNTOS DE TREINAMENTO E VALIDAÇÃO

Interrupção precoce é um método comum utilizado para evitar o supertreinamento de uma rede neural (HAYKIN, 1998). Nesta técnica, os dados disponíveis para o treinamento são divididos em conjuntos de treinamento e validação. Durante o treinamento, o desempenho de generalização é medido no conjunto de validação e o treinamento termina quando o desempenho da generalização começa a diminuir. Na utilização de redes neurais é comum separar os dados em conjuntos de treinamento, validação e teste ordenados pelo tempo (HAYKIN, 1998). A fim de gerar os conjuntos de treinamento e validação ampliados ordenados pelo tempo procede-se da seguinte forma:

- Divide-se a série temporal original em períodos de treinamento, validação e teste disjuntos em ordem cronológica;
- Para cada conjunto geram-se padrões normais;
- Para cada conjunto e para cada padrão normal disponível, geram-se n padrões aleatórios normais;
- Para cada conjunto e para cada padrão normal disponível, geram-se $n+1$ padrões aleatórios anômalos;
- Para cada conjunto cria-se um conjunto padrão ampliado incluindo os verdadeiros padrões normais, padrões normais e anômalos aleatórios.

O problema é que a divisão de dados em ordem cronológica faz com que informações importantes para o treinamento se percam. Assim, uma nova forma de divisão referida sob a designação de divisão distribuída é proposta por Oliveira (2004). Na

divisão distribuída à série temporal original é dividida em conjunto de treinamento, validação e teste, em ordem cronológica. No entanto, certo número de padrões aleatórios gerados a partir do conjunto de treino são adicionados para formar o conjunto de validação ampliado e vice-versa. Desta forma, os resultantes conjuntos de treinamento e validação ampliados terão informações de todo o período disponível para o treinamento. Isso pode resultar em um melhor desempenho de classificação para classificadores que necessitam de um conjunto de validação para o treinamento a fim de evitar um supertreinamento.

A geração de conjuntos de treinamento e de validação ampliados na divisão distribuída funciona da seguinte maneira:

- Divide-se a série temporal original em períodos de treinamento e teste disjuntos em ordem temporária;
- Para cada conjunto geram-se padrões normais;
- Para cada padrão normal disponível no conjunto de treinamento geram-se n padrões aleatórios normais. Seleciona-se uma percentagem destes padrões para formar o conjunto de treinamento ampliado e a percentagem restante para formar o conjunto de validação ampliado;
- Para cada padrão normal disponível no conjunto de treinamento geram-se $n+1$ padrões aleatórios anômalos. Seleciona-se uma percentagem desses padrões para formar o conjunto de treinamento ampliado e o restante para formar o conjunto de validação ampliado;

3.2.2.3 APRIMORANDO A RBF-DDA ATRAVÉS DA SELEÇÃO DE θ^-

Oliveira (2004) observou experimentalmente que para alguns problemas de classificação o valor de θ^- tem uma influência considerável sobre o desempenho do modelo RBF-DDA, e então propôs um método para aumentar seu desempenho através da seleção adequada do valor de θ^-.

Este valor influencia diretamente no número de unidades na camada intermediária adicionadas pelo DDA durante o treinamento, θ^- pequeno pode produzir um grande número de unidades ocultas, o que implica em um método que classifica corretamente todas as amostras de treinamento, porém, não possui uma boa generalização, podendo causar um supertreinamento. Para evitar este supertreinamento, utiliza-se uma parte dos dados de treinamento para formar um conjunto de validação, e o conjunto de validação para selecionar o valor de θ^- que atinge o melhor desempenho para o modelo.

Primeiramente, dividem-se os dados de entrada em conjuntos de treinamento (chamado de conjunto de treinamento completo) e conjunto de teste. Em seguida, divide-se o conjunto de treinamento completo em um conjunto de treinamento reduzido e um conjunto de validação. Realiza-se o treinamento da RBF em duas fases, a primeira utiliza o conjunto de treinamento reduzido para treinar a RBF-DDA e o conjunto de validação para avaliar o desempenho de generalização, esta fase é realizada com valores decrescentes de θ^-, a fim de selecionar um θ^-_{opt}, o valor de θ^- que (quase) otimiza o desempenho da rede no conjunto de validação, ou seja, o desempenho de generalização. Na segunda fase, o treinamento é realizado utilizando o conjunto de treinamento completo com $\theta^- = \theta^-_{opt}$. O valor θ^+ padrão 0.4, é usado em ambas as fases de formação. Finalmente, o desempenho de generalização da rede RBF-DDA resultante é avaliado através de um conjunto de teste disjunto.

Neste trabalho testam-se uma série de valores θ^- utilizando o conjunto de validação, começando com o valor padrão, $\theta^- = 0.1$. Em seguida, θ^- é reduzido em $\theta^- = \theta^- \times 10^{-1}$, pois observa-se que o desempenho não muda significativamente para valores intermediários de θ^-. Diminui-se o valor de θ^- até que o erro de validação comece a aumentar, uma vez que valores menores levam ao supertreinamento. O valor quase ótimo encontrado por este procedimento é utilizado para se treinar com o conjunto de treinamento completo.

É importante salientar que o método introduzido aqui mantém duas importantes

características RBF-DDA: (a) a natureza construtiva do algoritmo e (b) a utilização eficaz de todos os dados de treinamento para ajustar os parâmetros do modelo (rede RBF) (OLIVEIRA, 2004).

3.3 AVALIAÇÃO DA QUALIDADE DOS MODELOS PROBABILÍSTICOS

3.3.1 CURVA ROC - RELATIVE OPERATING CHARACTERISTIC

A avaliação da qualidade dos modelos de previsão ou classificação baseia-se na análise de matrizes de contingência através de uma curva conhecida como *Relative Operating Characteristic* (curva ROC) que permite uma melhor visualização do problema de avaliação. Para isso, calculam-se medidas de desempenho, taxa de acerto (H) e taxa de falso alarme (F), ou seja, medidas de verificação que incidem sobre a correspondência entre as previsões e observações, tanto de forma individual como coletiva (JOLLIFFE; STEPHENSON, 2003). “Uma maneira natural de apresentar as estatísticas para a avaliação de um modelo de classificação é por meio de uma tabulação cruzada entre a classe prevista pelo modelo e a classe real dos exemplos. Essa tabulação é conhecida como tabela de contingência (também chamada de matriz de confusão)” (PRATI; BATISTA; MONARD, 2008).

Denominam-se as duas classes deste estudo como classe dos dados normais ($C_n = 0$) e classe das anomalias ($C_a = 1$). Na Tabela 1 é apresentado um modelo de tabela de contingência.

TABELA 1: Modelo de Tabela de Contingência

Evento Previsto	Evento Observado		TOTAL
	1	0	
1	VA	FA	VA+FA
0	FC	VC	FC+VC
TOTAL	VA+FC	FA+VC	VA+FA+FC+VC

Quando um evento observado como anomalia é classificado ou previsto como anomalia, ele é denominado verdadeiro alerta (VA), quando um evento observado como normal é classificado ou previsto como anomalia, ele é denominado falso alerta (FA).

Bem como para o caso dos eventos classificados como normais, quando um evento observado normal é classificado ou previsto como normal, ele é denominado verdadeiro correto (VC), quando um evento observado anômalo é classificado como normal, ele é denominado falso correto (FC), como se pode observar na Tabela 1.

A taxa de acerto (H) pode ser definida por

$$H = \frac{VA}{VA + FC} \tag{15}$$

correspondendo a proporção de ocorrências que foram corretamente previstas, neste caso, a proporção de dados de vazão anômalos que foram devidamente apontados.

A taxa de falso alarme (F) pode ser definida por

$$F = \frac{FA}{FA + VC} \tag{16}$$

correspondendo a proporção de não ocorrências que foram incorretamente previstas, neste caso, a proporção de dados de vazão normais que foram incorretamente previstos como anômalos.

Como citado anteriormente, ROC é um gráfico de taxa de acerto (eixo Y) contra taxa de falso alarme (eixo X), a Figura 12 é uma curva ROC típica. A localização da curva em todo o quadrado unitário é determinada pela capacidade de discriminação intrínseca do sistema de previsão, e a localização dos pontos específicos sobre uma curva é fixada pelo limiar de decisão no qual está funcionando o sistema (JOLLIFFE; STEPHENSON, 2003).

Uma curva ROC empírica pode ser plotada através de resultados de previsões emitidas como probabilidades numéricas ou classificações verbais de risco através de um limiar de decisão das previsões, cada limiar gerando uma tabela de contingência, e consequentemente os valores de H e F. Todas as curvas ROC empíricas possuem a forma básica mostrada na Figura 12, necessariamente passando por $(0,0)$ e $(1,1)$ e em outros lugares interiores ao quadrado unitário.

Um modelo devidamente calibrado é representado por uma curva ROC que se

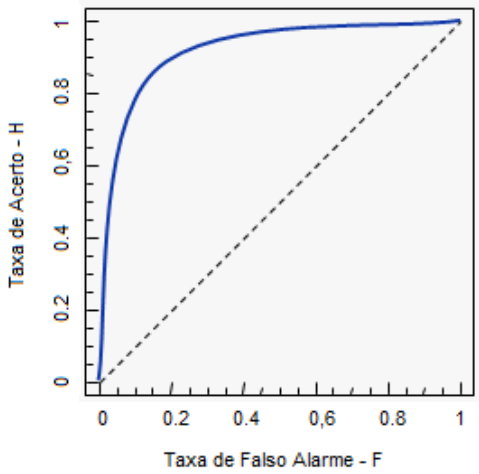


FIGURA 12: Curva ROC
FONTE: Adaptada de Jolliffe e Stephenson (2003)

eleva a partir de $(0,0)$ ao longo do eixo horizontal $(0,1)$, em seguida, se direciona para $(1,1)$. A diagonal $H = F$ representa habilidade zero do modelo, o que indica que as previsões são completamente não discriminatórias. Pontos abaixo da diagonal representam um sistema de previsão ou classificação mal calibrado indicando que previsões de anomalias não devem ser tomadas necessariamente como anomalias, e vice-versa.

A medida de acurácia de um modelo de previsão ou classificação mais utilizada é A_z, a área sob a curva ROC, também conhecida como AUC (*Area Under the Curve*), onde $A_z \in [0,1]$ e acurácia zero é indicada por $A_z = 0.5$, quando a curva se encontra ao longo da diagonal. $A_z = 1$ representa uma acurácia ótima, ou seja, a habilidade perfeita de um modelo ao realizar a previsão ou a classificação, e isto acontece quando a curva ROC sobe a partir de $(0,0)$ para $(0,1)$, e em seguida, para $(1,1)$. Um valor de A_z inferior a 0.5 corresponde a uma curva ROC abaixo da diagonal, que indica que o modelo está mal calibrado.

Neste trabalho calcula-se um indicador da presença de anomalias nos dados de vazão das estações hidrológicas estudadas através das probabilidades posteriores propostas pelo Teorema de Bayes que será apresentado na seção a seguir, e a partir destes indicadores encontram-se os limiares para a detecção de anomalias e

constroem-se as curvas ROC apresentadas no Capítulo 5 onde serão apresentados os resultados através da validação dos projetos construídos a partir de cada um dos modelos estudados.

3.4 GERAÇÃO DE PROBABILIDADES POSTERIORES CONDICIONADAS A UM INDICADOR - TEOREMA DE BAYES

3.4.1 TEOREMA DA PROBABILIDADE TOTAL

Esta seção baseia-se no livro de Theodoridis e Koutroumbas (2009) e destina-se à descrição de um indicador de classificação em um sistema de reconhecimento de padrões, baseando-se em argumentos probabilísticos decorrentes da natureza estatística dos recursos gerados. Opta-se por esta abordagem devido à variação estatística dos padrões de treinamento dos métodos propostos. Tendo como objetivo central projetar métodos que classificam um padrão desconhecido em uma classe mais provável, além de definir o que significa a expressão “mais provável”.

Dado um problema de classificação de M classes, $\omega_1, \omega_2, \ldots, \omega_M$, e um padrão desconhecido, que é representado por um vetor de características x, formam-se as M probabilidades condicionais $P(\omega_i|x), i = 1,2,\ldots,M$, que podem ser chamadas de probabilidades posteriores. Em outras palavras, cada uma destas parcelas representa a probabilidade de o padrão desconhecido pertencer à respectiva classe ω_i, uma vez que o vetor de características correspondente assume o valor x. Sendo assim estas probabilidades condicionais representam o termo “mais provável”. Assim, os classificadores calculam o máximo destes M valores ou, de forma equivalente, o máximo de uma função apropriada definida por eles. O padrão desconhecido é então designado para a classe correspondente a esse valor máximo.

Utiliza-se o teorema de Bayes para efetuar o cálculo das probabilidades condicionais, e de funções de densidade de probabilidade com base na evidência experimental disponível, ou seja, nos vetores de características que correspondem aos padrões do conjunto de treinamento.

3.4.1.1 TEORIA DA DECISÃO DE BAYES

Para este trabalho serão consideradas duas classes, a classe dos dados normais e a classe dos dados anômalos. Sejam ω_1 e ω_2 as duas classes que os padrões de treinamento pertencem, e $P(\omega_1)$, $P(\omega_2)$ as probabilidades a priori conhecidas, e se não forem conhecidas pode-se calculá-las a partir dos vetores de características de treinamento disponíveis. De fato, se N é o número total de padrões de treinamento disponíveis, e N_1, N_2 pertencem a ω_1 e ω_2, respectivamente, então $p(\omega_1) \approx N_1/N$ e $p(\omega_2) \approx N_2/N$.

As outras grandezas estatísticas que se presume serem conhecidas são as funções de densidade de probabilidade de classe condicional $p(x|\omega_i), i = 1,2,\ldots,M$ que descrevem a distribuição dos vetores característicos em cada uma das classes, e se forem desconhecidas, também podem ser estimadas a partir dos dados de treinamento disponíveis. Esta função $p(x|\omega_i)$ também é conhecida como a função de probabilidade condicional de x dado ω_i. Deve-se ressaltar o fato de que um pressuposto implícito foi feito, ou seja, os vetores característicos podem assumir qualquer valor no espaço de características multi-dimensional. No caso que vetores característicos assumem apenas valores discretos, funções de densidade $p(x|\omega_i)$ se tornam probabilidades e são denotadas por $P(x|\omega_i)$.

Assim pode-se efetuar o cálculo das probabilidades condicionais através do Teorema de Bayes (Apêndice A):

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)} \tag{17}$$

onde $p(x)$ é a função densidade de probabilidade de x e para o qual tem-se (Apêndice A):

$$p(x) = \sum_{i=1}^{2} p(x|\omega_i)P(\omega_i) \tag{18}$$

A regra de classificação de Bayes pode agora ser indicada como

$$\begin{aligned} P(\omega_1|x) > P(\omega_2|x),\ x \text{ é classificado em } \omega_1 \\ P(\omega_1|x) < P(\omega_2|x),\ x \text{ é classificado em } \omega_2 \end{aligned} \tag{19}$$

O caso de igualdade é prejudicial, pois o padrão pode ser atribuído a qualquer uma das duas classes. Utilizando a Equação 17 a decisão pode ser equivalentemente baseada nas desigualdades:

$$p(x|\omega_1)P(\omega_1) \gtrless (x|\omega_2)P(\omega_2) \tag{20}$$

$p(x)$ não é levada em conta, pois é a mesma para todas as classes e que não afeta a decisão. Além disso, se as probabilidades a priori são iguais, isto é, $P(\omega_1) = P(\omega_2) = 1/2$, a Equação 20 se torna:

$$p(x|\omega_1) \gtrless (x|\omega_2) \tag{21}$$

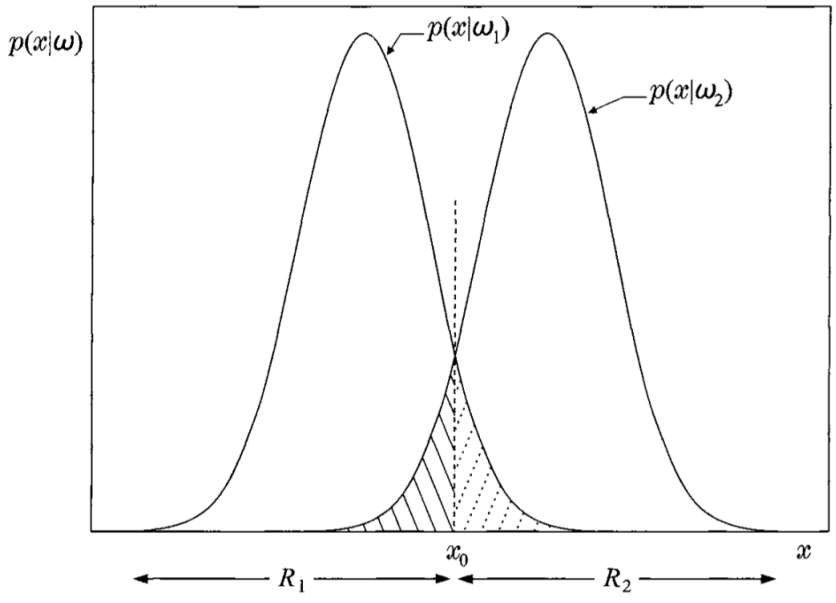


FIGURA 13: Regiões R_1 e R_2 formadas pelo classificador Bayesiano
FONTE: Theodoridis e Koutroumbas (2009)

Assim, os valores da função densidade de probabilidade condicional avaliada em x determinam a busca pelo valor máximo. A Figura 13 apresenta um exemplo de duas

classes equiprováveis e mostra as variações de $p(x|\omega_i), i = 1,2$, como funções de x para o caso simples de uma única característica $(l = 1)$. A linha tracejada em x_0 é um limite particionador do espaço de características em duas regiões, R_1 e R_2. De acordo com a regra de decisão de Bayes, para todos os valores de x em R_1 o classificador decide pela classe ω_1 e para todos os valores de x em R_2 decide pela classe ω_2. No entanto, é evidente a partir da figura que erros de decisão são inevitáveis. De fato, há uma probabilidade finita de um x pertencer à região R_2 e, ao mesmo tempo, pertencer à classe ω_1, e o mesmo é verdadeiro para os pontos provenientes da classe ω_2 que pertencem à região R_1. Assim a probabilidade total, P_e, de cometer um erro de decisão é dada por:

$$2P_e = \int_{-\infty}^{x_0} p(x|\omega_2)dx + \int_{x_0}^{+\infty} p(x|\omega_1)dx \tag{22}$$

que é igual à área total do sombreado sob as curvas na Figura 13.

O classificador Bayesiano é ótimo no que diz respeito a minimizar a probabilidade do erro de classificação, uma prova formal desta afirmação pode ser encontrada em Theodoridis e Koutroumbas (2009). Além disso, os autores afirmam que estes resultados podem ser generalizados para um problema de multi classes. Em um problema de classificação com M classes, $\omega_1, \omega_2, \ldots, \omega_M$, um padrão desconhecido, representado pelo vetor de características x, é classificado como membro da classe ω_i se:

$$P(\omega_i|x) > P(\omega_j|x) \qquad \forall j \neq i \tag{23}$$

E esta escolha também minimiza a probabilidade de erro de classificação (THEODORIDIS; KOUTROUMBAS, 2009).

4 PROJETOS DE APLICAÇÃO

A previsão pressupõe que o método realiza uma estimativa do valor real de vazão ou até mesmo do degrau de vazão. Métodos de previsão estão baseados na hipótese de que o valor a ser previsto de um dado é o valor real que deveria pertencer àquele dado. Portanto os mais diversos trabalhos de previsão utilizam-se de modelos lineares, regressivos, redes neurais, interpolações, entre outros métodos, a fim de realizar diferentes tipos de previsão de dados em uma série. Para que isto ocorra é necessária a construção de projetos de aplicação adequados aos diferentes métodos propostos.

Este capítulo destina-se a apresentação dos projetos de aplicação construídos para os métodos estudados. Além disso, examina um método para detecção de anomalias em um problema de classificação, a rede neural de função de base radial juntamente com o algoritmo de decaimento dinâmico (RBF-DDA). Idealmente, uma rede neural treinada deve classificar apenas padrões pertencentes às classes disponíveis durante as fases de treinamento, padrões aleatórios (aqui chamados de padrões anômalos) e padrões de classes não disponíveis durante o treinamento devem ser rejeitados pela rede. O principal objetivo é aplicar a RBF-DDA nos dados de vazão dos postos hidrológicos estudados, a fim de classificá-los individualmente e pontualmente como normais ou anômalos.

4.1 ABORDAGENS DE PREVISÃO

4.1.1 MÉTODO SOM

A técnica *Self - Organizing Maps* (SOM) é aplicada para prever elementos de uma amostra constituída por degraus de vazão, ou dados de vazão propriamente ditos. Sua saída retorna uma classificação dos dados de entrada e através desta classificação

obtém-se os valores da previsão para as amostras analisadas. Se a diferença entre o valor da previsão e do dado de vazão (ou degrau de vazão) original for significativa, então considera-se a amostra como sendo uma anomalia.

Segundo Valença (2005) os passos necessários à construção de um modelo de rede neural são baseados num processo iterativo que compreende as etapas de: análise preliminar e identificação do modelo; treinamento da rede; e verificação da rede (etapa de teste).

O autor ressalta que a definição da topologia da rede que envolve fatores como a definição do conjunto de treinamento, a complexidade do fenômeno que se está analisando, a relação e a quantidade de neurônios de entrada e saída, é uma tarefa aparentemente complicada, já que geralmente é feita de maneira empírica e não existem regras que as determinem, mas ao mesmo tempo é uma tarefa que influenciará significativamente o desempenho do modelo. Com isso, para que exista uma otimização da arquitetura da rede neural, necessita-se da avaliação de um grande número delas, com diferentes parâmetros, ou seja, é necessário que se treine diversas configurações até que se encontre a melhor arquitetura para determinado problema em análise.

Além disso, o treinamento da rede neural depende do ajuste de seus parâmetros assim, para o problema de detecção de anomalias em dados de vazão, no qual se dispõe de padrões de entrada, mas não se conhece previamente os padrões de saída para o treinamento, opta-se por utilizar um treinamento não supervisionado.

Para a verificação da rede, utiliza-se de um conjunto de testes, ou seja, conjunto de dados diferente do conjunto utilizado para o treinamento, de modo que o conjunto de teste seja composto por uma amostra representativa dos dados do problema em estudo. Caso o resultado obtido não seja satisfatório, escolhe-se uma nova topologia para a rede, e repete-se o ciclo.

Após os períodos de treinamento e de teste, considera-se que a rede está pronta para ser utilizada, consistindo em uma poderosa e ágil ferramenta, pois todas as infor-

mações da amostra utilizada com o treinamento ficaram armazenadas nos neurônios após o ajuste de seus pesos.

Nas seções seguintes serão explanadas as metodologias utilizadas para o período de treinamento e teste do SOM.

4.1.1.1 TREINAMENTO

Para prever os dados de vazão dos postos hidrológicos de Porto Amazonas (PA) e União da Vitória (UV) no estado do Paraná, e classificá-los como corretos ou anômalos, utiliza-se o SOM de forma inovadora, no sentido de que será feita a previsão de degraus horários de vazão considerando-se um horizonte passado e futuro de 6h com relação ao dado a ser analisado, e a partir desta previsão o dado intermediário de uma amostra de 13 elementos será classificado como correto, ou anômalo.

• Dados de Entrada:

Selecionou-se os dados de entrada para o treinamento da rede neural a partir dos dados de vazão $(q_1, q_2, ..., q_n)$ dos postos hidrológicos de UV e PA dispostos em intervalos de uma hora, pertencentes aos anos de 1998 até 2007, totalizando 87.648 dados pertencentes a um período de 10 anos. Estes dados passaram previamente pela consistência realizada por técnicos capacitados pelo SIMEPAR, ou seja, o período de treinamento, em princípio, não possui intervalos de falhas, nem dados inconsistentes.

Freire (2009) propôs em seu trabalho, com o objetivo de aplicar redes neurais em previsões hidrológicas probabilísticas, um método que não considera simplesmente os valores de vazão previstos, e sim os valores de degraus de vazão (d_q), que são, na verdade, a diferença entre a vazão em um momento futuro e a última vazão observada (Q_{obs}), considerando um intervalo de 6 horas. Segundo a autora, é necessário que se forme um vetor de doze valores de degraus de vazão subsequentes que represente o comportamento da vazão ao longo de 72 horas. Identificando um conjunto de comportamentos de vazão em 72 horas que representam, com um erro pequeno, o histórico de vazões observadas que contemplam as mais diversas situações hidrológicas, pois

assim a previsão pode ser apresentada na forma desse comportamento sem perdas significativas nos valores de vazão informados, e com grande ganho na simplificação do método.

Portanto, no intuito de classificar dados individualmente acrescenta-se o degrau, ao qual o dado a ser investigado pertence, entre os 12 degraus propostos por Freire (2009), formando um vetor de 13 valores de degraus subsequentes de vazão o que representará uma amostra para este estudo, onde o dado a ser analisado é o sétimo elemento da amostra e consideram-se os seis dados anteriores e posteriores a ele para que se efetue a previsão. Porém, a fim de obter-se uma previsão mais realista neste trabalho utilizam-se degraus de vazão subsequentes de 1 hora, contemplando um horizonte de 13 horas, enquanto Freire (2009), que visava outros objetivos de pesquisa, utilizou degraus de 6 horas e contemplou um horizonte de 72 horas.

Assim, necessita-se dispor os dados de vazão de UV e PA em degraus de vazão subsequentes, de 1h em 1h, $\Delta q_i = q_{i+1} - q_i$, onde $i = 1, ..., n-1$, formando uma matriz A_{SOM} (EQUAÇÃO 24) de dados de entrada, de dimensão $m \times 13$, em que cada uma de suas linhas representa uma amostra da população dos dados analisados, e a sétima coluna da matriz representa os dados a serem investigados. Assim, centraliza-se esta pesquisa na investigação dos dados da sétima coluna desta matriz, levando em consideração a relação deles com os degraus das 6 colunas anteriores, e 6 colunas posteriores.

$$A_{SOM} = \begin{bmatrix} q_2 - q_1 & q_3 - q_2 & \cdots & q_8 - q_7 & \cdots & q_{13} - q_{12} & q_{14} - q_{13} \\ q_3 - q_2 & q_4 - q_3 & \cdots & q_9 - q_8 & \cdots & q_{14} - q_{13} & q_{15} - q_{14} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ q_m - q_{m-1} & q_{m+1} - q_m & \cdots & q_{m+6} - q_{m+5} & \cdots & q_{m+11} - q_{m+10} & q_{m+12} - q_{m+11} \end{bmatrix} \quad (24)$$

Como o número de dados de vazão dos postos hidrológicos estudados são extremamente extensos, para processamento e análise das metodologias propostas fez-se uso da linguagem e ambiente para computação estatística R Core Team (2012), juntamente com seus diversos pacotes.

• Parâmetros:

Aplica-se o SOM do pacote *kohonen* (WEHRENS; BUYDENS, 2007) do software R Core Team (2012) com a utilização dos seguintes parâmetros:

1. O conjunto de dados completo foi apresentado 100 vezes à rede em cada rodada;

2. A taxa de aprendizagem (α) está diretamente relacionada com o incremento dos pesos durante o processo de treinamento. Valores altos aceleram o treinamento, entretanto podem causar instabilidade e saturação. Por outro lado, valores baixos podem tornar o treinamento lento (VALENÇA, 2005). Geralmente estes valores estão situados entre $0 < \alpha < 1$. Para esta pesquisa a taxa α utilizada diminui linearmente de 0.05 até 0.01, a cada iteração;

3. O raio da vizinhança começa com um valor que abrange $2/3$ de todas as distâncias de unidade para unidade e diminui a cada iteração;

4. Os representantes iniciais apresentados à rede, ou seja, os valores iniciais dos pesos são gerados aleatoriamente a partir dos dados de entrada, de acordo com o número de *codebooks*, neste trabalho utiliza-se a palavra *codebook* a fim de representar os neurônios da saída da rede neural, além disso, eles representam amostras que possuem características comuns entre si e, portanto possuem o mesmo formato dessas amostras, ou seja, são vetores de dimensão 13, onde o sétimo elemento é o representante dos elementos das amostras a serem analisadas. Assim, se os dados de entrada são alocados em n *codebooks*, então serão sorteadas n amostras dos dados de entrada para comporem os valores iniciais dos pesos.

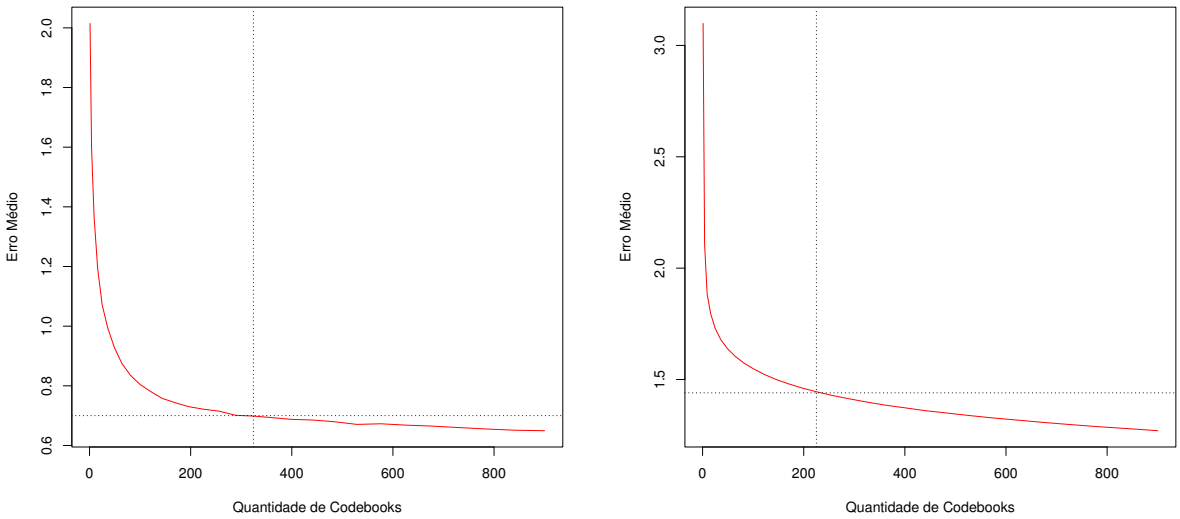
5. Utilizaram-se grades do tipo hexagonal neste trabalho.

Testaram-se 30 tipos diferentes de grades hexagonais $j \times j$ em que $j = 1, 2, 3, \ldots, 30$ para cada estação hidrológica, executando 10 rodadas para cada configuração, e totalizando 300 rodadas. Selecionaram-se, entre estas, as rodadas que resultaram as

menores médias das distâncias entre as amostras e os *codebooks* que as representam, a fim de analisá-las com o intuito de determinar o número de *codebooks* a ser utilizado.

• Escolha do Número de *Codebooks*

A partir das diversas simulações, das 10 rodadas de cada configuração das grades, calculou-se um erro médio considerando a raiz quadrada de cada uma das 300 menores médias divididas por 13 (dimensão das amostras), e através dos resultados obtidos, ilustrados na Figura 14, e a partir da estabilidade encontrada, definiu-se 324 como sendo a quantidade ideal de *codebooks* para os dados da estação de Porto Amazonas, formando uma grade hexagonal 18×18, e 225 como sendo a quantidade ideal de *codebooks* para os dados do problema da estação de União da Vitória, formando uma grade hexagonal 15×15.



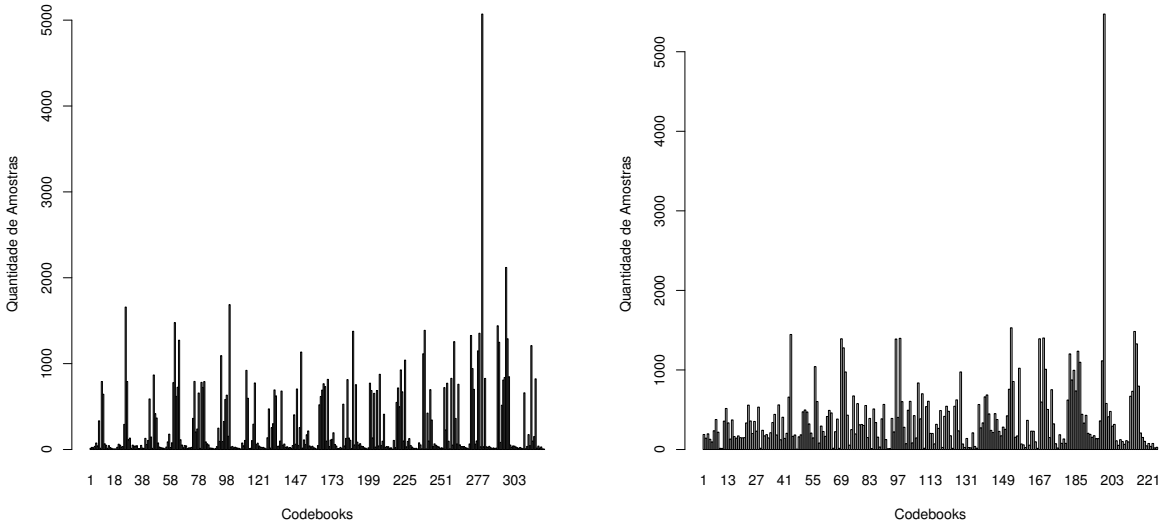
(A) Porto Amazonas (B) União da Vitória

FIGURA 14: Número de *codebooks*: (A) Porto Amazonas e (B) União da Vitória

FONTE: A autora (2014)

Com isso, definiu-se que cada *codebook* representa a classificação de um determinado número de amostras semelhantes entre si. Os gráficos da Figura 15 ilustram a organização destas distribuições, apresentando a quantidade de amostras classificadas em cada um dos *codebooks* para as duas estações hidrológicas estudadas, indicando que existem *codebooks* que representam uma grande quantidade de amos-

tras, enquanto que outros representam um número muito pequeno delas.



(A) Porto Amazonas (B) União da Vitória

FIGURA 15: Número de amostras representadas por cada *codebook*: (A) Porto Amazonas e (B) União da Vitória

FONTE: A autora (2014)

4.1.1.2 PERÍODO DE TESTE - RECONHECIMENTO DE PADRÕES

Posterior à fase de treinamento dos dados através do SOM, inicia-se a fase de teste da rede neural, ou seja, a busca propriamente dita de anomalias em dados de vazão não consistidos de um período posterior ao período de treinamento. Segundo Valença (2005), a avaliação do desempenho de um modelo deve ser realizada, sobre o conjunto de verificação. Desta forma, construiu-se uma matriz semelhante à matriz dos dados de entrada, porém, com os degraus de vazão originais, construídos a partir de dados não consistidos, dos anos de 2008 a 2010 posteriores ao período de treinamento, dos postos hidrológicos de Porto Amazonas e União da Vitória.

Através da função *map* do pacote *kohonen* (WEHRENS; BUYDENS, 2007) do software R (R Core Team, 2012), e do resultado do treinamento realizado anteriormente, classificou-se as amostras da nova matriz do período de teste de acordo com a numeração dos *codebooks* resultantes do período de treinamento. Além disso, aplicou-se a função *map* às amostras, com degraus originais, retirando-se os seus sétimos ele-

mentos a fim de realizar a previsão dos valores relativos a estes elementos em cada uma das amostras. A partir desta previsão calculou-se o quadrado da diferença entre o sétimo valor das amostras originais pertencentes ao período de teste (valores observados) e o sétimo valor dos *codebooks* representantes destas amostras (valores previstos), e por meio destas diferenças obteve-se um indicador da presença de anomalias nos dados de vazão, de forma que quanto mais significativa a diferença entre a previsão e o dado de vazão original, maior a chance de considerar este dado como sendo um dado anômalo, ou seja, esta diferença representa a indicação de força de determinado dado ser anômalo ou não. Por meio desta diferença conclui-se a aplicação do SOM para os degraus de vazão.

Os resultados desta metodologia, criada para a aplicação do SOM, aos dados de vazão das estações hidrológicas de Porto Amazonas e União da Vitória serão apresentados no Capítulo 5 juntamente com os demais resultados obtidos através das aplicações dos métodos de interpolação *Smooth Spline* e das Redes Neurais RBF que serão apresentados a seguir.

4.1.2 MÉTODO *SMOOTH SPLINE*

Neste trabalho aplicam-se *Smooth Splines* com o mesmo objetivo que aplica-se o SOM, ou seja, para prever os valores dos sétimos elementos de uma amostra constituída por valores de vazão, a fim de detectar anomalias nos dados das estações hidrológicas de Porto Amazonas e União da Vitória.

No projeto de aplicação desenvolvido para a utilização do *Smooth Spline*, a técnica foi aplicada diretamente sobre dados originais (não consistidos) de vazão (não sobre degraus de vazão consistidos). Esta é uma vantagem deste modelo em relação aos demais, pois enquanto o SOM, apresentado anteriormente, e a RBF, que será explanada a seguir, necessitam de um período de dados consistidos, ou seja, dados que passaram por uma análise prévia que eliminou todas as anomalias da sequência, para fazer o treinamento das redes, o *Smooth Spline* classifica os dados diretamente a

partir dos dados originais, não consistidos, e não transformados em degraus, apesar de que os valores dos dados de vazão consistidos ainda serão necessários para a avaliação do projeto de aplicação construído, como será apresentado adiante.

Para prever os dados de vazão dos postos hidrológicos de Porto Amazonas (PA) e União da Vitória (UV) no estado do Paraná, e classificá-los como corretos ou anomalias, utiliza-se o *Smooth Spline* de maneira semelhante ao SOM, no sentido de que faz-se a previsão de valores horários de vazão considerando-se um horizonte de 6h passadas e 6h futuras com relação ao dado a ser analisado, e a partir desta previsão o dado intermediário de uma amostra de 13 elementos será classificado como correto, ou anômalo.

Sendo assim, para que os resultados dos métodos possam ser comparados no Capítulo 5, selecionou-se os dados a partir dos valores de vazão originais $(q_1, q_2, ..., q_n)$ dos postos hidrológicos de PA e UV dispostos em intervalos de uma hora, pertencentes aos anos de 1998 até 2007, totalizando 87.648 dados pertencentes a um período de 10 anos, representando um período de "treinamento", e os dados dos próximos 3 anos de 2008 até 2010 representando o período de "teste", apesar desta separação não ser necessária. Estes dados são considerados brutos obtidos através do banco de dados do SIMEPAR, ou seja, que devem conter anomalias.

De maneira análoga ao SOM, constrói-se uma matriz A_{spline} (EQUAÇÃO 25) de 13 colunas de valores de vazão em que cada uma de suas linhas representam uma amostra a ser estudada, e para cada amostra faz-se a previsão de seu sétimo elemento considerando-se o comportamento dos 6 elementos passados e dos 6 futuros. Porém neste método não foram considerados degraus de vazão e sim, a diferença de todos os valores de vazão de uma amostra com o primeiro valor desta amostra, ou seja, no lugar de degraus de vazão utiliza-se uma normalização destes valores através

destas diferenças com relação aos primeiros elementos de cada amostra.

$$A_{spline} = \begin{bmatrix} q_1 - q_1 & q_2 - q_1 & \cdots & q_7 - q_1 & \cdots & q_{12} - q_1 & q_{13} - q_1 \\ q_2 - q_2 & q_3 - q_2 & \cdots & q_8 - q_2 & \cdots & q_{13} - q_2 & q_{14} - q_2 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ q_m - q_m & q_{m+1} - q_m & \cdots & q_{m+6} - q_m & \cdots & q_{m+11} - q_m & q_{m+12} - q_m \end{bmatrix} \quad (25)$$

Novamente, para este método, fez-se uso da linguagem e ambiente para computação estatística R Core Team (2012), por meio da função *smooth.spline* que ajusta uma função *Smooth Spline* cúbica aos dados fornecidos através da matriz de 13 colunas (visando a predição da sétima coluna) construída com os dados de vazão normalizados dos períodos de "treinamento" e de "teste", que seguem a mesma metodologia, porém foram considerados separadamente para simplificação da apresentação dos resultados no Capítulo 5, e como auxílio na escolha do parâmetro de suavização.

Na função *smooth.spline* deve-se definir o valor do parâmetro de suavização previamente, para este fim realizou-se testes empíricos: no período de treinamento variou-se o valor do parâmetro como apresentado na Tabela 3 do Anexo A, e para cada um dos resultados obtidos construiu-se uma curva ROC e calculou-se o valor de sua área (descritos no Capítulo 3, e complementados no Capítulo 5), realizou-se uma comparação entre os valores de AUC obtidos a fim de encontrar o melhor valor para o parâmetro de suavidade de cada posto hidrológico estudado.

Para a construção da curva ROC, também foi necessário utilizar a série consistida equivalente ao período de treinamento. Porém, se a escolha do parâmetro de suavização não fosse necessária então a construção de uma série consistida também não seria. Isto pode acontecer, por exemplo, com as próximas aplicações do *Smooth Spline* para as estações de PA e UV (que, a partir de agora, possuem parâmetro definido) ou também, para estações com comportamentos parecidos aos delas.

Novamente recorrendo à Tabela 3 do Anexo A conclui-se que os melhores valores para estes parâmetros são 0.50 e 0.75 para as estações de Porto Amazonas e União da Vitória respectivamente.

Posteriormente à aplicação da função *smooth.spline* aos dados de treinamento das duas estações estudadas, e da definição dos coeficientes de suavidade, aplicou-se a função *smooth.spline* aos dados de teste, transformados em matriz da mesma maneira que os dados de treinamento (A_{spline}), e retirando-se os sétimos elementos de cada amostra a fim de realizar a previsão de seus valores.

Como consequência obteve-se uma matriz, de mesma dimensão da matriz de entrada, com os resultados da interpolação e previsão resultantes desta aplicação, inclusive para os sétimos elementos de cada amostra. A partir disso calculou-se, novamente, o quadrado da diferença entre o sétimo valor das amostras originais (valores observados) e o sétimo valor das amostras obtidas através da interpolação por *Smooth Spline* (valores previstos), e a partir destas diferenças obteve-se um indicador da presença de anomalias em dados de vazão, de forma que, novamente, quanto mais significativa a diferença entre a previsão e o dado de vazão original, maior a chance do dado considerado ser uma anomalia, ou seja, o valor desta diferença indica quão anômalo o dado pode ser.

Os resultados da aplicação desta metodologia, criada para a aplicação dos *Smooth Splines*, aos dados de vazão das estações hidrológicas de União da Vitória e Porto Amazonas serão apresentados no Capítulo 5 juntamente com os demais resultados obtidos através das aplicações dos métodos de Redes Neurais SOM e RBF.

4.2 ABORDAGEM DE CLASSIFICAÇÃO

4.2.1 MÉTODO RBF - DDA

Neste trabalho utiliza-se a rede neural RBF juntamente com o algoritmo DDA a fim de realizar a classificação dos dados de vazão das estações hidrológicas de Porto Amazonas e União da Vitória no sentido de detectar anomalias. Como a saída da rede é formada por variáveis contínuas utilizadas para classificação, espera-se que o método retorne um indicador da possibilidade de cada valor da série de dados ser, ou não, uma anomalia. Diferentemente do SOM e do *Smooth Spline* o método não

retorna uma previsão dos dados de vazão, ou dos degraus de vazão propriamente ditos.

Oliveira (2004) aplicou esta mesma abordagem de classificação com objetivos diferentes: obtendo resultados para janelas de séries temporais. Uma janela de série temporal é formada por w pontos de dados consecutivos extraídos de uma série temporal. Em contrapartida, o objetivo desta pesquisa é testar o método RBF-DDA a fim de analisar se esta técnica permite a avaliação da presença de anomalias pontuais em cada dado da série temporal de vazão.

Conforme a teoria apresentada no Capítulo 3 a RBF-DDA possui dois parâmetros cruciais para o seu treinamento os limites θ^+ e θ^-. No âmbito desta aplicação utiliza-se o valor padrão $\theta^+ = 0.4$, e varia-se θ^- (de maneira similar à proposta de Oliveira (2004)) a fim de encontrar um valor ótimo (θ^-_{opt}). Além disso, utiliza-se uma das técnicas de treinamento negativo proposto, ou seja, gera-se uma amostragem negativa sintética, anomalias dos diversos tipos, tamanhos, em diversas situações, com diferentes durações, sempre de acordo com os dados observados, para que a amostra possua a mesma quantidade de anomalias e não anomalias dobrando o período de treinamento para que a rede esteja preparada para detectar casos esporádicos, pois os casos anômalos, individualmente, não seguem regras podendo ocorrer a qualquer momento. E por fim, a separação dos dados em período de treinamento completo, abrangendo o período de treinamento e o período de validação, além do período de teste que também faz parte deste método.

Para classificar os dados de vazão dos postos hidrológicos de Porto Amazonas (PA) e União da Vitória (UV) no estado do Paraná como corretos ou anômalos, utiliza-se a RBF-DDA de maneira semelhante ao SOM e ao *Smooth Spline*, no sentido de que faz-se a classificação dos valores dos degraus de vazão considerando-se um horizonte de 6h passadas e 6h futuras com relação ao degrau a ser analisado, e a partir disso o degrau intermediário de uma amostra de 13 elementos será classificado como correto, ou anômalo.

Sendo assim, para que os resultados dos métodos possam ser comparados no Capítulo 5, novamente selecionou-se os dados a partir dos valores de vazão consistidos $(q_1, q_2, ..., q_n)$ dos postos hidrológicos de UV e PA dispostos em intervalos de uma hora, pertencentes aos anos de 1998 até 2007, totalizando 10 anos, para formar o período de treinamento completo. E a partir dos valores de vazão originais pertencentes aos anos de 2008 a 2010, totalizando 3 anos, forma-se o período de teste da rede neural RBF-DDA.

De maneira análoga ao SOM, constrói-se uma matriz A_{RBF} (EQUAÇÃO 26) de 13 colunas de degraus de vazão em que cada uma de suas linhas representam uma amostra a ser estudada.

$$A_{RBF} = \begin{bmatrix} q_2 - q_1 & q_3 - q_2 & \cdots & q_8 - q_7 & \cdots & q_{13} - q_{12} & q_{14} - q_{13} \\ q_3 - q_2 & q_4 - q_3 & \cdots & q_9 - q_8 & \cdots & q_{14} - q_{13} & q_{15} - q_{14} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ q_m - q_{m-1} & q_{m+1} - q_m & \cdots & q_{m+6} - q_{m+5} & \cdots & q_{m+11} - q_{m+10} & q_{m+12} - q_{m+11} \end{bmatrix} \quad (26)$$

Novamente, para este método, fez-se uso da linguagem e ambiente para computação estatística R Core Team (2012), por meio do pacote RSNNS (BERGMEIR; BENÍTEZ, 2012) e das diversas funções que envolvem redes RBF e o algoritmo DDA. Através destas funções e da matriz A_{RBF} (EQUAÇÃO 26) formada pelos dados de treinamento completo, diversos tipos de anomalias (que obedecem as características das anomalias já observadas em dados não consistidos) foram introduzidas às amostras, para que existam a mesma quantidade de dados normais e dados anômalos. A partir deste conjunto ampliado formou-se o período de treinamento completo ampliado da RBF-DDA, e dividiu-se este período em período de treinamento (equivalente a 70% dos dados de treinamento completo ampliado) e período de validação (30% restantes dos dados de treinamento completo ampliado).

Na função do pacote RSNNS do software R que executa o algoritmo DDA, utiliza-se, para esta abordagem, o parâmetro $\theta^+ = 0.4$ e varia-se θ^- decrescendo seu valor a partir do padrão 0.1, treina-se a rede com o período de treinamento e com cada um

dos θ^- variados a fim de maximizar o valor de AUC (Capítulos 3 e 5) encontrados através da aplicação do resultado da rede aplicado ao período de validação. Com base nestes testes e nos resultados dos cálculos das AUC apresentados na Tabela 4 do Anexo A, conclui-se que os melhores valores para estes parâmetros são 1×10^{-2} e 5×10^{-5} para as estações de Porto Amazonas e União da Vitória respectivamente.

Posteriormente à utilização das funções contidas no pacote RSNNS para aplicar o algoritmo DDA e treinar a rede RBF com dados de vazão das estações de PA e UV, com os parâmetros θ^-_{opt} definidos para cada um dos casos, e os conjuntos de treinamento, validação e teste devidamente discriminados e ampliados através da técnica de amostragem negativa, obteve-se a saída da rede representada por duas unidades (neurônios), formando uma matriz de tantas linhas quanto a quantidade de entradas da rede e duas colunas contendo a medida da possibilidade de uma entrada ser ou não uma anomalia, sendo que a primeira coluna representa a classe dos dados normais na qual quanto maior o valor do resultado da amostra mais normal ela é considerada pela rede, e a segunda coluna representa a classe dos dados anômalos em que quanto maior o valor do resultado da amostra mais anômala ela é considerada pela rede.

Para a RBF-DDA, diferentemente do SOM e *Smooth Spline*, a saída representa diretamente o indicador da presença de anomalias em dados de vazão, já que não retorna uma previsão do dado analisado e sim sua classificação direta, com isso, neste método, não há necessidade do cálculo do quadrado da diferença entre dado previsto e observado como nos métodos apresentados anteriormente.

Os resultados desta metodologia, criada para a aplicação da RBF-DDA, aos dados de vazão das estações hidrológicas de Porto Amazonas e União da Vitória serão apresentados no Capítulo 5 juntamente com os demais resultados obtidos através das aplicações dos métodos de Redes Neurais SOM e interpolação *Smooth Spline*.

5 VALIDAÇÃO DOS PROJETOS

Neste capítulo propõe-se um método de avaliação da acurácia e validação do desempenho dos modelos propostos para previsão dos dados de vazão de ambas as estações hidrológicas analisadas, SOM e *Smooth Spline*, e também do modelo de classificação dos dados em normais ou anômalos através da RBF-DDA (CAPÍTULOS 3 e 4).

5.1 PREVISÃO

Como apresentado no Capítulo 4 os métodos utilizados para previsão dos dados, SOM e *Smooth Spline*, retornam o valor previsto do degrau de vazão e do dado analisado propriamente dito, e a partir destes valores calcula-se o quadrado da diferença entre o valor previsto pelos métodos e o valor observado proveniente dos dados originais a fim de encontrar indicadores de anomalias (dif). Através do Teorema de Bayes (CAPÍTULO 3) e da normalização[1] de dif , transformam-se os indicadores de anomalias em probabilidades dos dados serem anômalos ou não, ou seja, encontra-se uma medida da tendência do dado ser ou não uma anomalia. Com isso, constroem-se curvas ROC (CAPÍTULO 3) para os períodos de treinamento e teste de cada um dos métodos propostos e através do cálculo da área sob as curvas (AUC) conclui-se se os métodos estão aptos a serem utilizados para o objetivo deste estudo e até mesmo decide-se qual é o melhor modelo a ser utilizado.

Para a construção da curva ROC (CAPÍTULO 3) utilizam-se funções do pacote ROCR (SING *et al.*, 2009) do software R Core Team (2012). Como esta técnica realiza avaliação, necessita-se comparar as probabilidades resultantes da aplicação do

[1] A normalização é feita através da função tRank do pacote multic (LUNDE *et al.*, 2013) do software R Core Team (2012) a fim de transformar dif utilizando uma distribuição normal de quantis empíricos.

Teorema de Bayes com valores de referência.

Estes valores de referência são representados por um vetor binário (ww) com coordenadas iguais a 1 onde há a ocorrência de uma anomalia e coordenadas nulas onde os dados são considerados normais. Esta atribuição é obtida através da diferença relativa entre dados consistidos[2] e os dados originais do mesmo período de estudo (anos de 1998 a 2007 para o período de treinamento e 2008 a 2010 para o teste). Assim, se esta diferença for maior do que uma certa tolerância relativa definida aqui como 2% (valor que elimina pequenas anomalias não significativas à este estudo) considera-se a presença de uma anomalia, e valores menores que esta tolerância representam dados normais.

Para a aplicação do Teorema de Bayes a partir da Equação 17, considera-se:

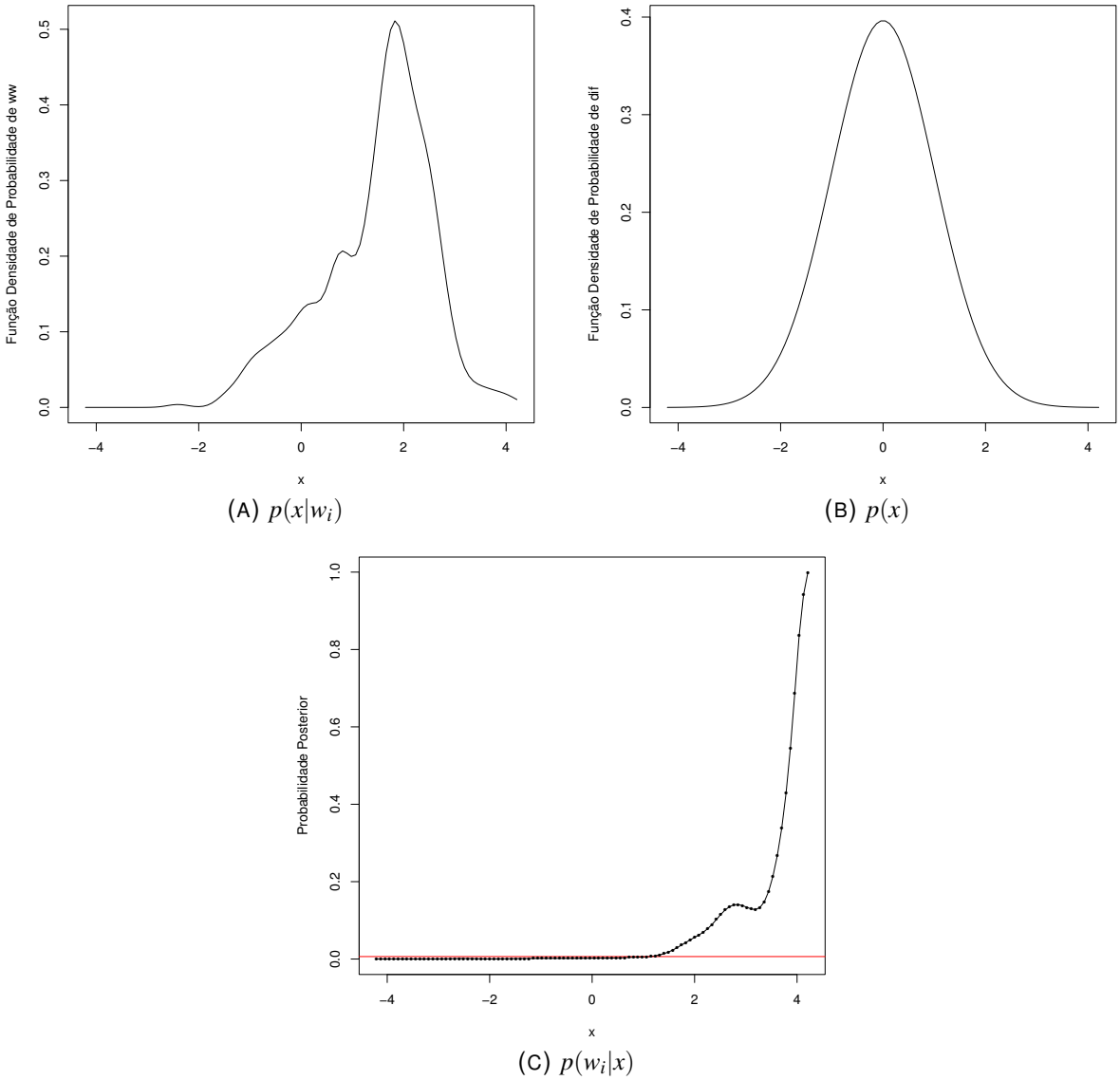
- $p(x|w_i)$, com $w_i = ww$ e $x = dif$, como sendo a função densidade de probabilidade de ww, ou de x condicional a w_i, ou seja, a probabilidade do método utilizado indicar que o dado é uma anomalia sabendo que uma anomalia foi realmente observada através de ww;
- $P(w_i)$ como a probabilidade a priori de ocorrência de anomalia calculada através de ww;
- $p(x)$ a função densidade de probabilidade incondicional de dif, ou seja, do método utilizado indicar que o dado é uma anomalia;

Com estes parâmetros calcula-se a probabilidade posterior ($p(w_i|x)$) as probabilidades dos dados de entrada serem anômalos, ou não, sabendo que o método utilizado o apontou como anômalo através de dif. E assim, constrói-se a curva ROC através da comparação entre $p(w_i|x)$ e ww.

[2]Dados que passaram por uma correção manual dos diferentes tipos de anomalias, *spikes*, oscilações diárias, ruídos, falhas, ou extremos de mudanças de *offset*, encontrados em dados de vazão nos períodos estudados, realizada previamente por técnicos capacitados do SIMEPAR.

5.1.1 SOM

Através do treinamento do SOM com os dados de vazão das estações hidrológicas de Porto Amazonas (PA) e União da Vitória (UV), realizado no período de 10 anos de 1998 a 2007 (descrito no Capítulo 4) e da aplicação do Teorema de Bayes (descrito anteriormente) obtém-se os gráficos de $p(x|w_i)$, $p(x)$ e $p(w_i|x)$ (FIGURA 16) referentes à PA e os mesmos gráficos referentes à UV (FIGURA 17).



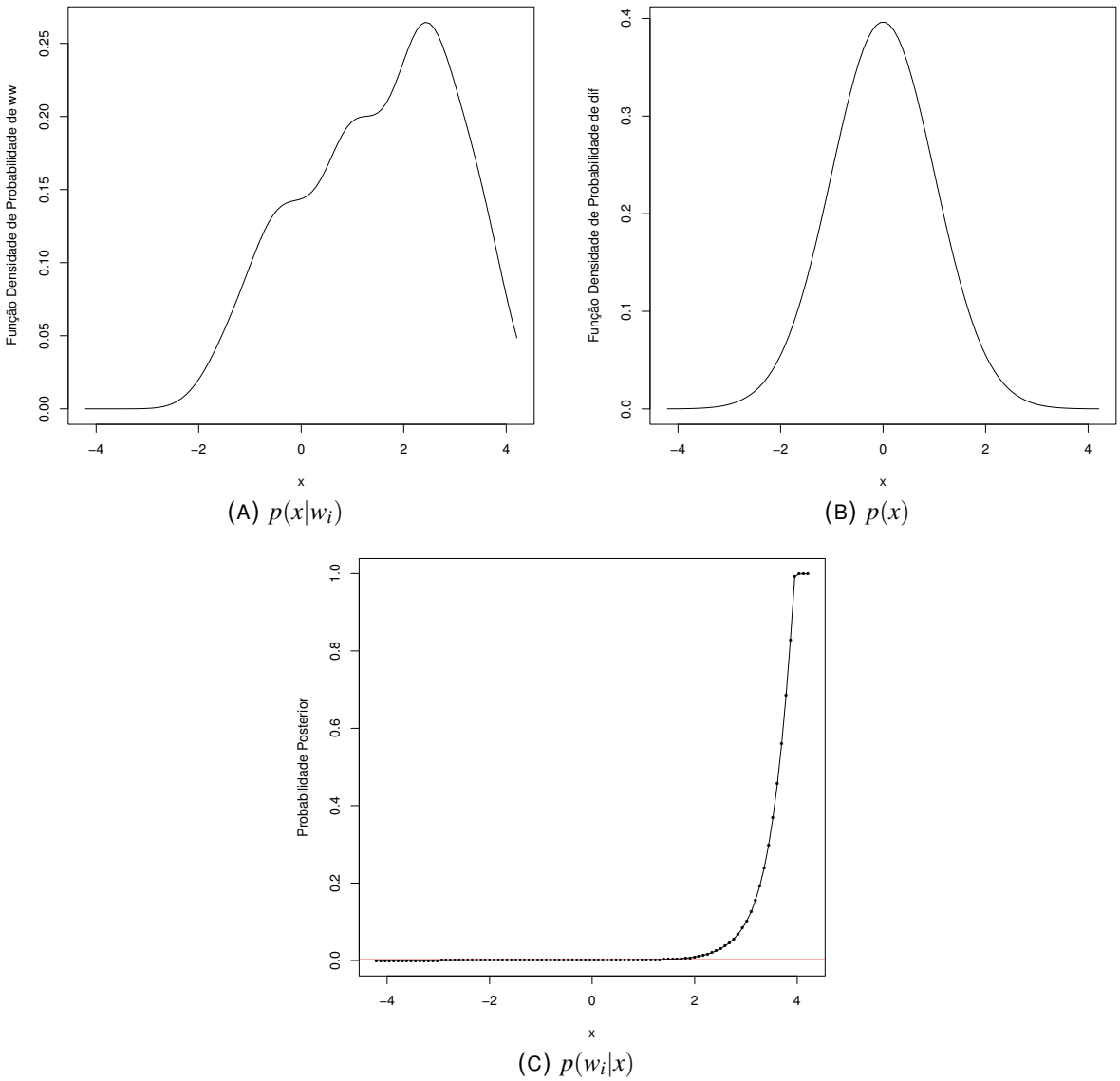
(A) $p(x|w_i)$ (B) $p(x)$

(C) $p(w_i|x)$

FIGURA 16: Parâmetros do Teorema de Bayes para o SOM - Porto Amazonas: (A) $p(x|w_i)$, (B) $p(x)$ e (C) $p(w_i|x)$

FONTE: A autora (2014)

E através dos resultados de $p(w_i|x)$, constrói-se a curva ROC e obtém-se o valor



(A) $p(x|w_i)$ (B) $p(x)$ (C) $p(w_i|x)$

FIGURA 17: Parâmetros do Teorema de Bayes para o SOM - União da Vitória: (A) $p(x|w_i)$, (B) $p(x)$ e (C) $p(w_i|x)$

FONTE: A autora (2014)

da área sob a curva de 0.844 para PA e 0.798 para UV, resultados satisfatórios, pois estão significativamente acima de 0.5. O fato do valor de UV ser ligeiramente menor que o valor encontrado para PA deve estar associado à uma maior diversidade de situações hidrológicas em UV, pois observou-se através de ww que PA possui uma maior quantidade de anomalias do que UV fazendo com que a rede neural SOM tenha uma maior facilidade em seu processo de aprendizado.

Para o período de teste, envolvendo os 3 anos (2008 a 2010) , por meio da apli-

cação do resultado obtido através do Teorema de Bayes e da construção da curva ROC calculam-se valores, esperadamente inferiores com relação ao período de treinamento, para a área sob a curva de 0.841 para PA e 0.748 para UV, indicando que o método SOM cumpre seu papel na detecção de anomalias em dados de vazão. A comparação destes resultados com os outros métodos propostos será apresentada das seções seguintes.

5.1.2 *SMOOTH SPLINE*

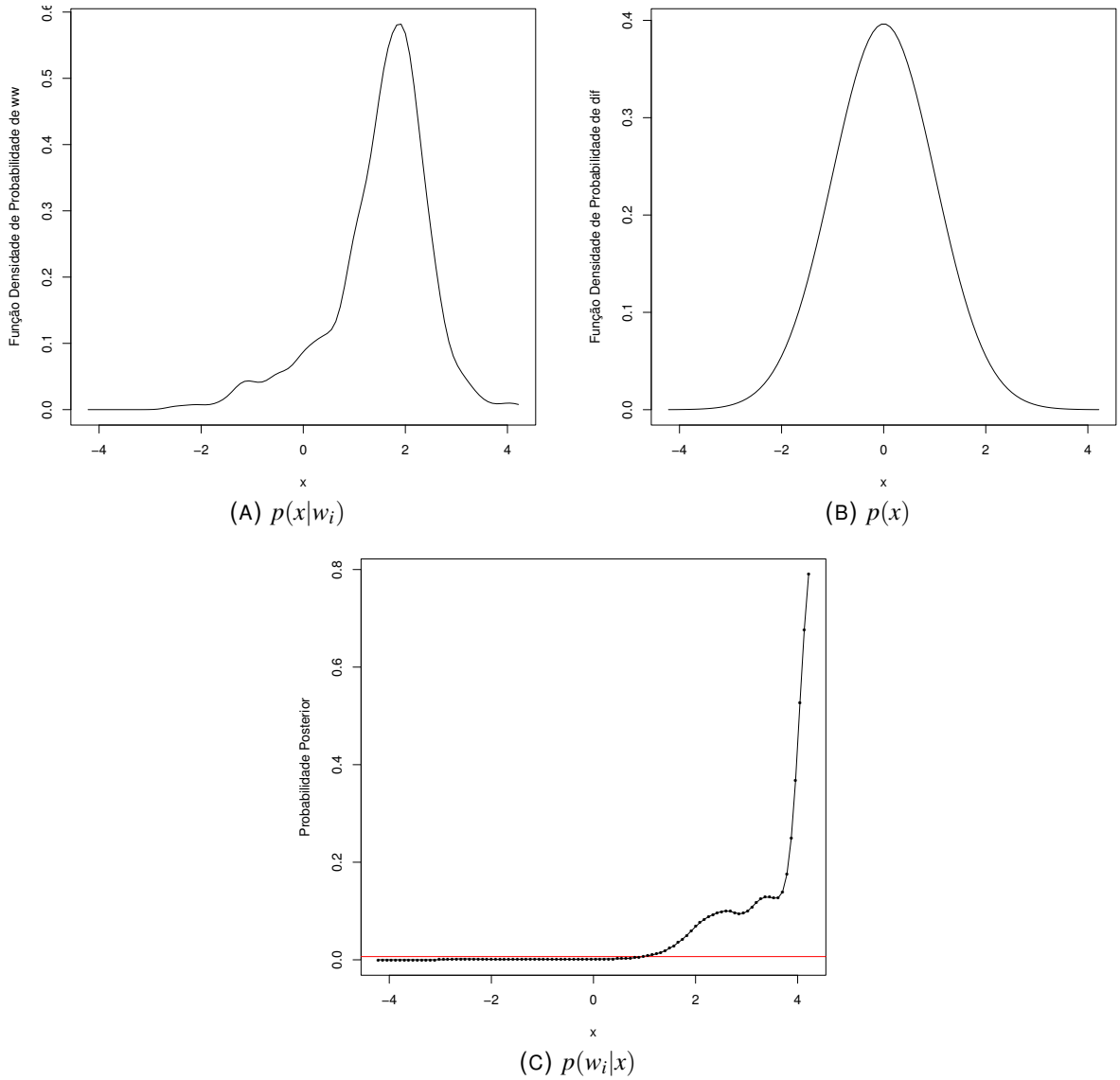
A técnica *Smooth Spline* recebeu avaliação semelhante ao SOM. Apesar de, como dito no Capítulo 3, este método dispensar a separação do conjunto de dados em períodos de treinamento e teste, esta discriminação foi utilizada a cargo de comparação entre os dois métodos propostos para previsão dos dados e também para comparação entre eles e o método de classificação RBF-DDA.

Para o período de treinamento (anos de 1998 a 2007) obtêm-se, por meio da aplicação do Teorema de Bayes os gráficos de $p(x|w_i)$, $p(x)$ e $p(w_i|x)$ (FIGURA 18) referentes à PA e os mesmos gráficos referentes à UV (FIGURA 19).

A área sob a curva ROC estimada para o período de treinamento (anos de 1998 a 2007) na estação de Porto Amazonas é de 0.872, e para União da Vitória de 0.849, satisfazendo as mesmas considerações levantadas na análise do SOM, porém com valores ligeiramente maiores.

Para o período de teste (2008 a 2010) a área sob a curva é estimada em 0.875 para PA e 0.845 para UV, indicando que o *Smooth Spline* cumpre, ainda mais que o SOM, seu papel na detecção de anomalias em dados de vazão, com o diferencial de que dispensando a distinção dos períodos de treinamento e teste, os valores da curva ROC não mudam significativamente entre estes períodos a partir da aplicação do *Smooth Spline*.

A comparação destes resultados com os outros métodos propostos será apresentada das seções seguintes.



(A) $p(x|w_i)$ (B) $p(x)$

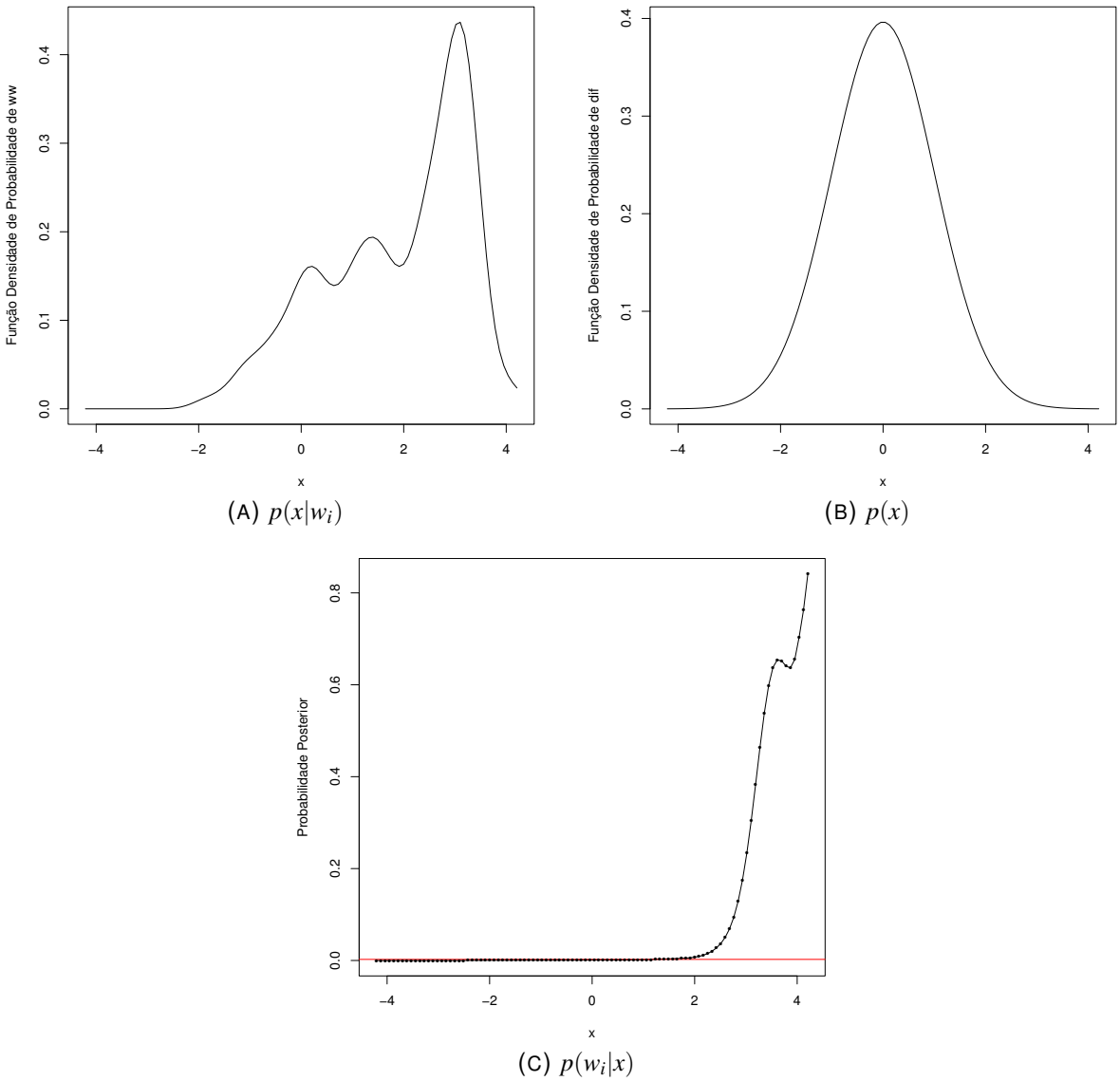
(C) $p(w_i|x)$

FIGURA 18: Parâmetros do Teorema de Bayes para o *Smooth Spline* - Porto Amazonas: (A) $p(x|w_i)$, (B) $p(x)$ e (C) $p(w_i|x)$

FONTE: A autora (2014)

5.2 CLASSIFICAÇÃO

Diferentemente dos métodos de previsão apresentados anteriormente, a RBF-DDA, utilizada como método de classificação, retorna como saída um vetor com valores indicadores da intensidade de um dado de vazão ser ou não anômalo, correspondente ao dif dos métodos apresentados anteriormente, porém substituindo a previsão realizada por uma medida própria de tendência. E semelhantemente ao SOM e *Smooth Spline* transforma-se estes indicadores de anomalias em probabilidades



(A) $p(x|w_i)$ (B) $p(x)$ (C) $p(w_i|x)$

FIGURA 19: Parâmetros do Teorema de Bayes para o *Smooth Spline* - União da Vitória: (A) $p(x|w_i)$, (B) $p(x)$ e (C) $p(w_i|x)$

FONTE: A autora (2014)

por meio da aplicação do Teorema de Bayes (CAPÍTULO 3), para que através dele, construam-se as curvas ROC[3] para os períodos de treinamento e teste da RBF-DDA, calculem-se as áreas sob estas curvas, e comparem-se estes resultados com os demais métodos.

Semelhantemente à técnica de avaliação utilizada para a validação dos métodos SOM e *Smooth Spline* construiu-se um vetor binário ww de referência para ambos

[3]Novamente utilizando as funções do pacote ROCR (SING *et al.*, 2009) do software R Core Team (2012)

os períodos (teste e treinamento) considerando-se os diferentes possíveis tipos de anomalias a serem encontrados em dados de vazão.

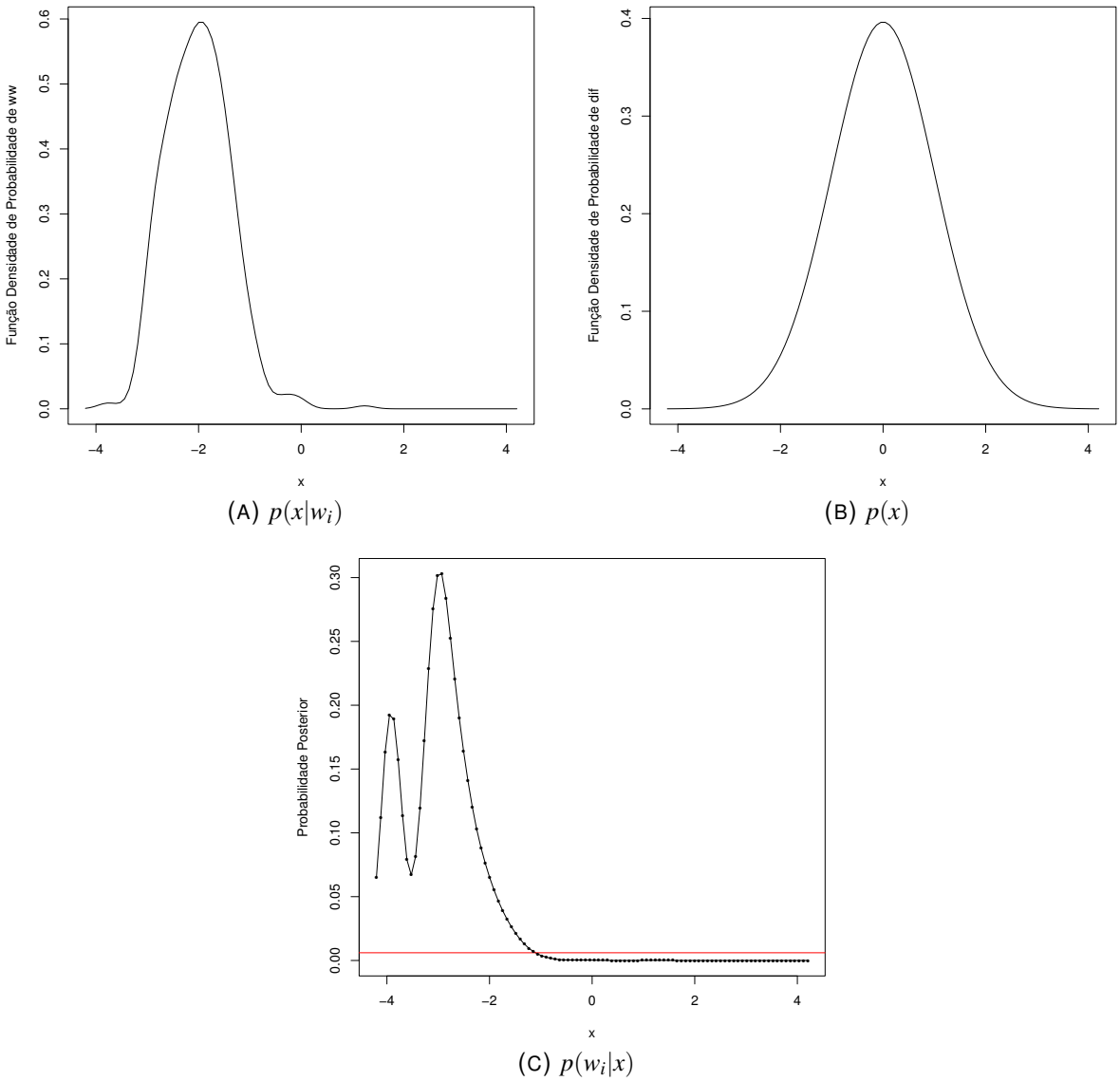
5.2.1 RBF-DDA

Como citado anteriormente, a rede neural RBF juntamente com o algoritmo de aperfeiçoamento DDA foram aplicados aos dados de vazão das estações hidrológicas de Porto Amazonas e União da Vitória a fim de realizar uma classificação no sentido de apontar a intensidade de um dado ser ou não anômalo. Sendo assim, a saída da rede retorna os valores destes indicadores de anomalias. Aplica-se o Teorema de Bayes aos indicadores resultantes do período de treinamento da rede, e através de seus resultados transformam-se os indicadores de anomalias do período de teste em probabilidades dos dados serem ou não anômalos.

Através da utilização de ww e da aplicação da inferência bayesiana às saídas do treinamento, pode-se observar os resultados através dos gráficos de $p(x|w_i)$, $p(x)$ e $p(w_i|x)$ (FIGURA 20) referentes à PA e os mesmos gráficos referentes à UV (FIGURA 21).

A partir da utilização dos resultados de $p(w_i|x)$ para o período de treinamento, obtém-se o valor da área sob a curva de 0.958 para PA e 0.896 para UV. Aplicando os resultados das probabilidades condicionais obtidas aos resultados do período de teste as áreas sob a curva possuem valores de 0.902 para PA e 0.85 para UV, demonstrando a superioridade deste método para com os outros.

A próxima seção apresentará a comparação entre os diferentes métodos apresentados a fim de avaliar seus desempenhos na detecção de anomalias em dados de vazão das bacias hidrológicas estudadas.



(A) $p(x|w_i)$ (B) $p(x)$

(C) $p(w_i|x)$

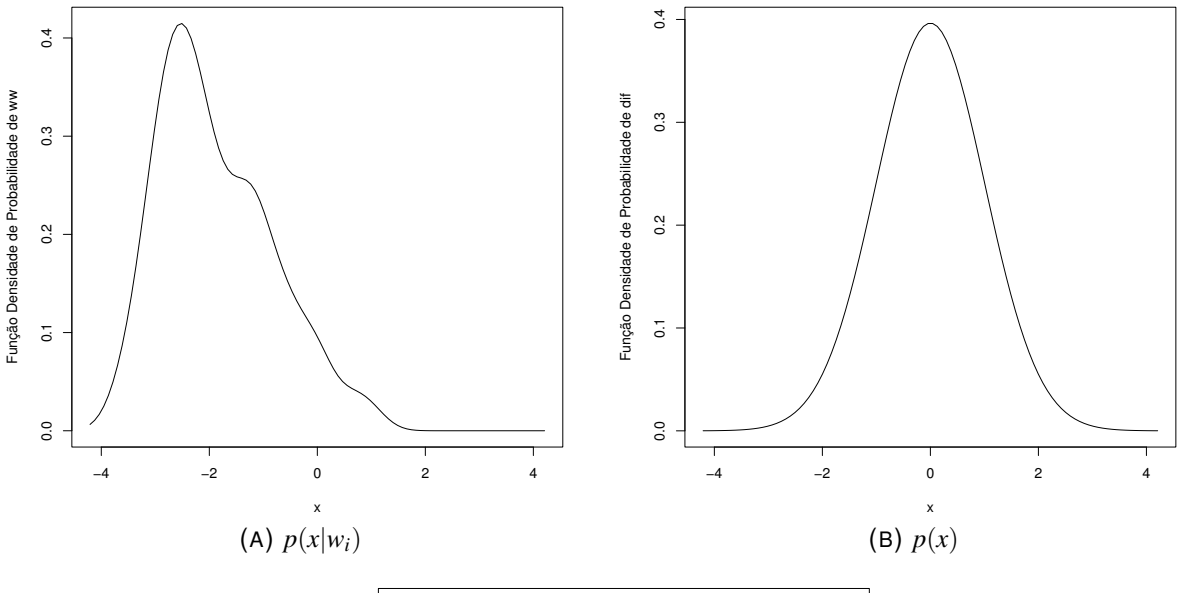
FIGURA 20: Parâmetros do Teorema de Bayes para a RBF-DDA - Porto Amazonas: (A) $p(x|w_i)$, (B) $p(x)$ e (C) $p(w_i|x)$

FONTE: A autora (2014)

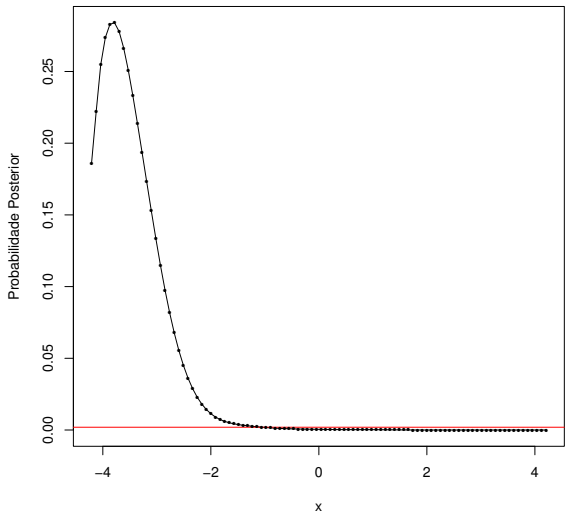
5.3 COMPARAÇÕES E RESULTADOS

A fim de cumprir os objetivos desta pesquisa aplicaram-se dois métodos de previsão, SOM e *Smooth Spline*, e um método de classificação, RBF-DDA, à detecção de anomalias em dados de vazão das estações hidrológicas de Porto Amazonas e União da Vitória, sub-bacias da bacia do rio Iguaçu.

A Tabela 2 apresenta a comparação dos resultados dos cálculos das áreas (AUC)



(A) $p(x|w_i)$ (B) $p(x)$



(C) $p(w_i|x)$

FIGURA 21: Parâmetros do Teorema de Bayes para a RBF-DDA - União da Vitória: (A) $p(x|w_i)$, (B) $p(x)$ e (C) $p(w_i|x)$

FONTE: A autora (2014)

das curvas ROC dos períodos de treinamento para os três métodos para ambas as estações estudadas.

TABELA 2: Valores de AUC - Período de Treinamento

Métodos \ Postos Hidrológicos	Porto Amazonas	União da Vitória
SOM	0.844	0.798
Smooth Spline	0.872	0.849
RBF-DDA	0.958	0.896

Os valores da Tabela 2 demonstram a superioridade do método RBF-DDA para o período de treinamento com relação aos demais métodos, os valores de AUC para ambas as estações estudadas estão, para todos os métodos apresentados, consideravelmente acima do valor crítico 0.5, porém, para o modelo desenvolvido para a aplicação do método RBF-DDA estes valores chegam próximos do valor ótimo 1.

Pode-se, além disso, observar que os resultados da estação de Porto Amazonas são relativamente maiores que os resultados de União da Vitória, pois o primeiro posto hidrológico é correspondente justamente a uma sub-bacia de cabeceira da bacia do rio Iguaçu, menor, e muito menos comportada do que as demais sub-bacias, implicando em uma maior ocorrência de diferentes tipos de dados anômalos, facilitando a aprendizagem dos métodos utilizados e sua aplicação à detecção de anomalias. E em comparação, a estação de União da Vitória possui características menos relevantes aos apontamentos de anomalias, devido ao fato de ser uma sub-bacia intermediária, central à bacia do rio Iguaçu, influenciada por diversas outras sub-bacias, e com um número consideravelmente menor de anomalias.

Apesar da avaliação do período de treinamento ser muito importante, no sentido de evidenciar se o método está realmente correspondendo às expectativas, é no período de teste que se realiza a validação dos projetos, pois este período apresenta aos modelos dados totalmente desconhecidos e diferenciados, portanto é no período de teste que pode-se realmente concluir se um determinado modelo está realmente correspondendo às expectativas com relação às novidades apresentadas.

As Figuras 22 e 23 apresentam as curvas ROC, bem como os valores de suas áreas, obtidas para os períodos de teste para os três métodos aplicados às estações de Porto Amazonas e União da Vitória respectivamente.

Estas curvas evidenciam a superioridade do método RBF-DDA, com relação ao SOM e ao *Smooth Spline*, para o período de teste de PA e UV. Deixando claro que para a sub-bacia de Porto Amazonas, que pode ser considerada uma sub-bacia de resposta rápida com relação à precipitação, menos comportada e com um número

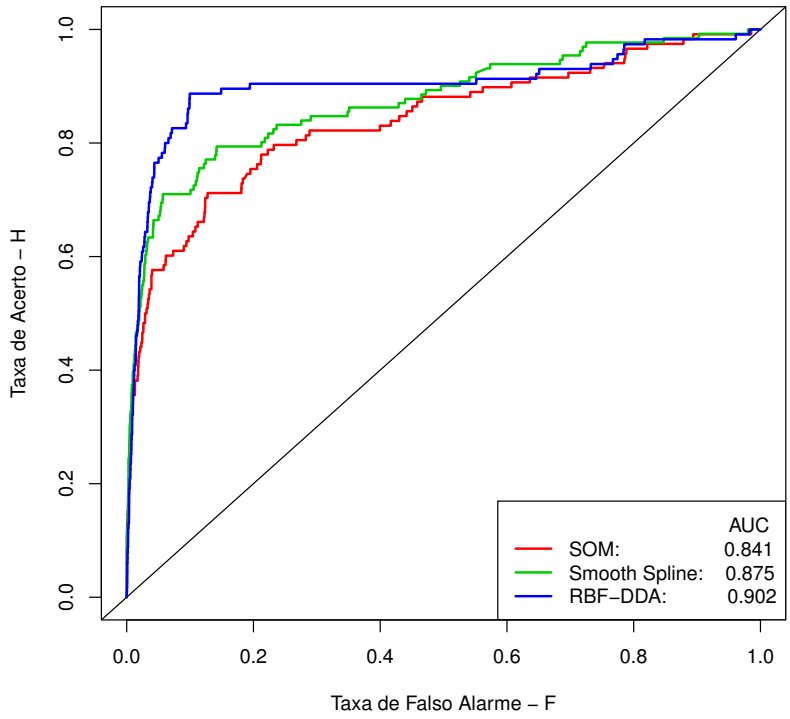


FIGURA 22: Curvas ROC dos diferentes métodos - Porto Amazonas

FONTE: A autora (2014)

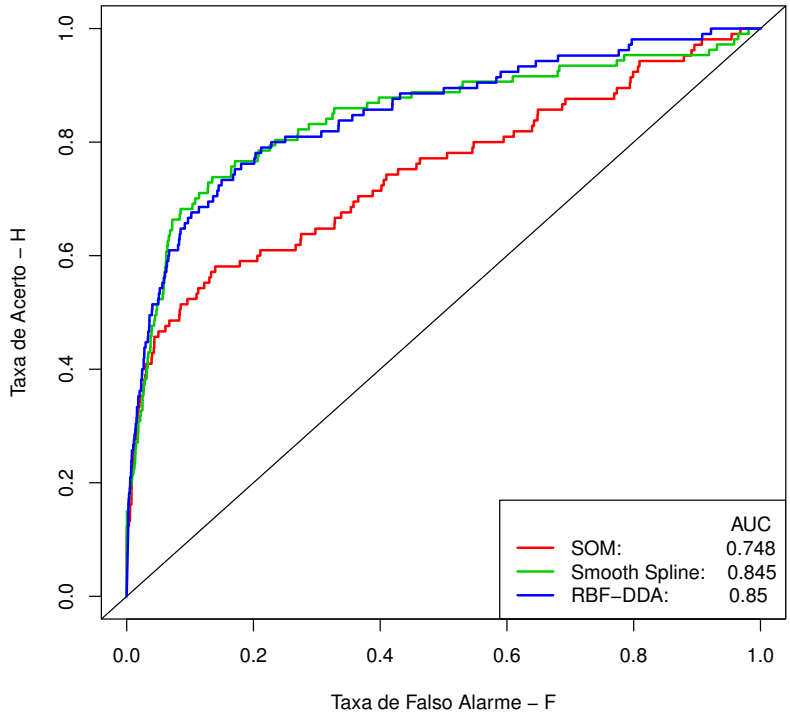


FIGURA 23: Curvas ROC dos diferentes métodos - União da Vitória

FONTE: A autora (2014)

maior de evidências anômalas o projeto de aplicação para RBF-DDA retornou resultados superiores com relação aos demais métodos. Em contrapartida para a sub-bacia de União da Vitória, que pode ser considerada uma sub-bacia de resposta lenta com relação à precipitação, mais comportada e com um número menor de anomalias, os projetos de aplicação para a RBF-DDA e o *Smooth Spline*, apresentaram resultados bastante similares. Nesta sub-bacia o método *Smooth Spline* se destacou, pois promove a interpolação dos dados, e como ficou comprovado, esses dados possuem pouca variação e poucas anomalias, com isso pode-se realizar uma interpolação muito mais correta.

5.4 AVALIAÇÃO VISUAL

Avaliações visuais podem ser feitas através de hidrógrafas (representações visuais do comportamento da vazão), compostas por dados de vazão consistidos, dados originais (brutos, extraídos diretamente do banco de dados do SIMEPAR), e pelos apontamentos de anomalias resultantes da aplicação dos diferentes métodos propostos.

Com o propósito de ilustrar os apontamentos das anomalias das estações de União da Vitória, têm-se as seguintes hidrógrafas oriundas das aplicações dos métodos de previsão SOM (FIGURA 24) e *Smooth Spline* (FIGURA 25) e para o método de classificação RBF-DDA (FIGURA 26).

Estas figuras evidenciam na cor vermelha dados com anomalias originais do banco de dados do SIMEPAR, em verde os dados que passaram pela técnica de consistência utilizada atualmente, e em preto os apontamentos sinalizados pelos modelos utilizados. Para o período ($22/02/2009$ a $27/02/2009$) da estação de União da Vitória evidenciou-se que as mais diversas falhas foram notadas e devidamente apontadas como anomalias pelos métodos utilizados através do SOM, *Smooth Spline* e RBF-DDA. Contudo, pode-se notar que existem dados normais apontados como anômalos indevidamente, bem como dados anômalos que não foram devidamente apontados

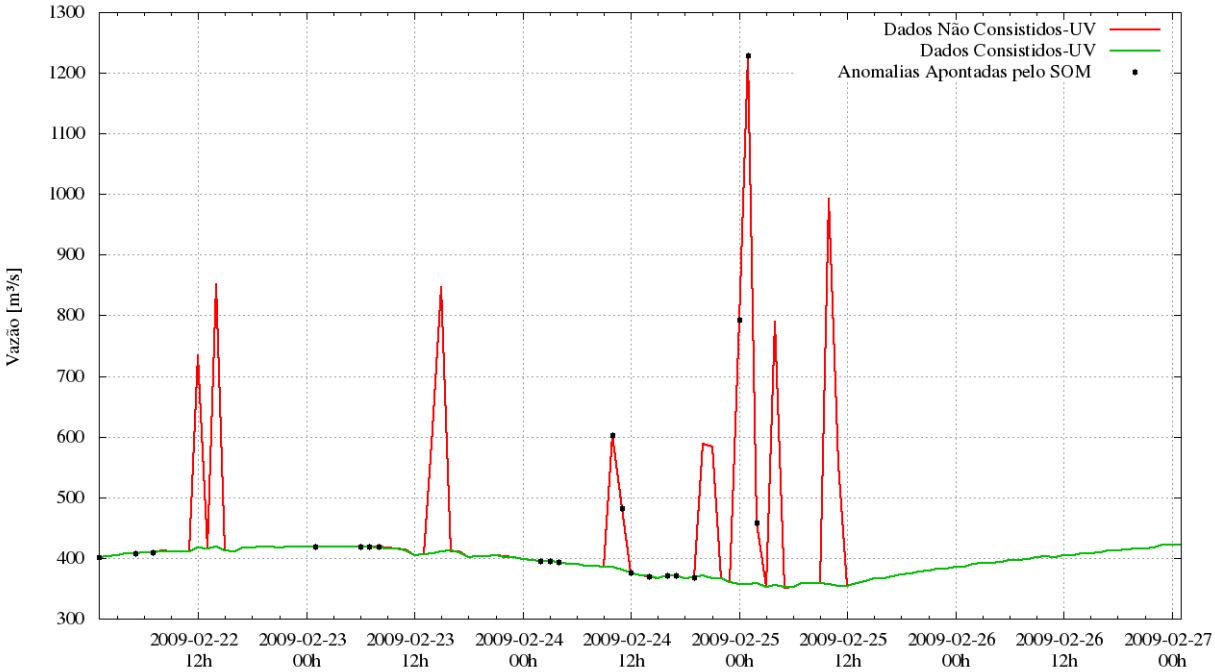


FIGURA 24: Dados apontados como anomalias através da técnica SOM - União da Vitória

FONTE: A autora (2014)

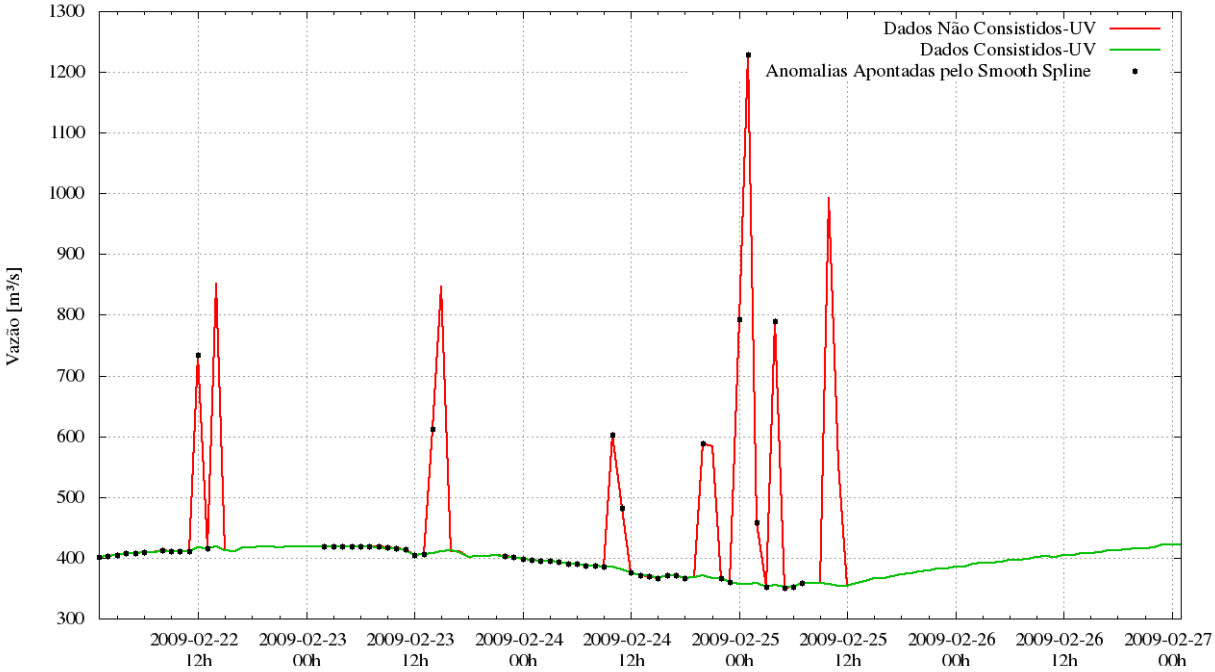


FIGURA 25: Dados apontados como anomalias através da técnica *Smooth Spline* - União da Vitória

FONTE: A autora (2014)

através das técnicas utilizadas, este fato justifica o valor das áreas da curva ROC não serem ótimos. Apesar disso, pode-se notar por meio das figuras que a rede RBF-DDA

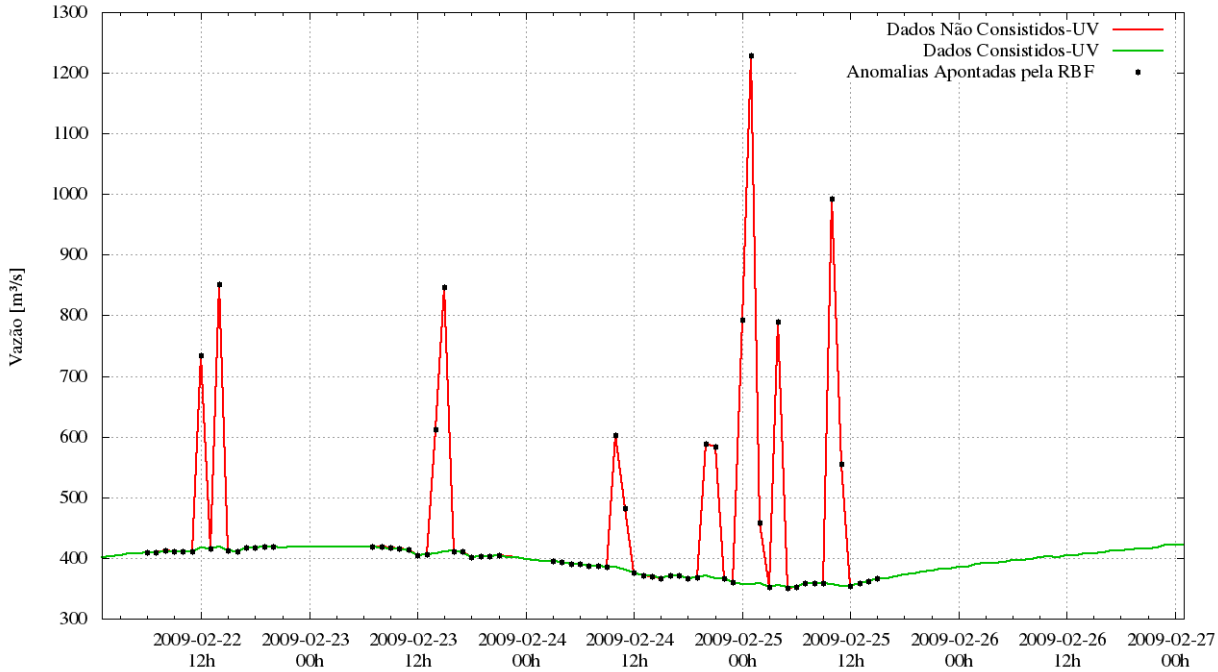


FIGURA 26: Dados apontados como anomalias através da técnica RBF-DDA - União da Vitória

FONTE: A autora (2014)

é superior aos demais métodos com relação aos apontamentos.

Com isso, podem-se considerar os métodos como sistemas de apoio ou suporte à decisão na identificação de dados anômalos, pois apontam dados de vazão como sendo dados anômalos para que técnicos capacitados decidam posteriormente, através de visualização gráfica, se estes dados são realmente anomalias, para que então a correção de períodos anômalos seja realizada, e as séries horárias de vazão se tornem totalmente consistidas e úteis aos sistemas de previsão de vazão utilizados pelo Instituto Tecnológico SIMEPAR, bem como em outras aplicações de planejamentos energéticos e projetos de empreendimentos hidráulicos.

6 CONCLUSÃO

A busca por padrões de comportamentos em dados de vazão torna-se uma ferramenta útil e muito necessária, principalmente em operações de reservatórios hidrológicos. Com isso, o objetivo principal deste trabalho foi desenvolver metodologias que detectassem dados anômalos em séries de vazão dos postos hidrológicos de União da Vitória e Porto Amazonas no estado do Paraná, a fim de facilitar o trabalho de consistência de séries de dados de vazão. Para tanto, fez-se um levantamento de padrões característicos de comportamento e principalmente os diferentes tipos de anomalias presentes em séries de dados de vazão.

Em comparação com dados previamente consistidos, em que se tem o conhecimento das diferentes características das anomalias encontradas em uma série de dados de vazão, os métodos: SOM, *Smooth Spline* e RBF-DDA, retornaram resultados satisfatórios, pois apontaram possíveis anomalias exatamente em áreas nas quais dados foram corrigidos quando passaram por uma consistência manual. Além disso, anomalias com todas as possíveis características, *spikes*, mudanças de *offset*, ruídos, oscilações diárias, entre outras, foram apontadas pelos métodos.

Técnicas de avaliação e visualização dos resultados através de curvas ROC, cálculos de suas áreas AUC, e representações gráficas a partir das hidrógrafas, permitiram a realização de comparação entre os métodos propostos, tanto nos períodos de treinamento quanto nos períodos de teste.

Através deste trabalho pode-se concluir que a técnica SOM de treinamento não supervisionado, abordada de maneira diferenciada, para realizar previsões de degraus de vazão, a fim de compará-las com dados existentes, e através destas comparações concluir se o dado possui comportamento normal ou anômalo corresponde às expectativas e cumpre os objetivos, porém com menor desempenho se comparada à

técnica de interpolação *Smooth Spline* que foi utilizada de forma semelhante ao SOM para realizar previsões mas que possui desempenho consideravelmente maior. Além disso, conclui-se que a técnica RBF-DDA empregada com fins classificatórios retornou os melhores resultados quando comparada com o SOM e o *Smooth Spline*, apresentando altos valores de AUC quando foi aplicada a duas sub-bacias (Porto Amazonas e União da Vitória) de comportamentos tão diferentes.

Em trabalhos futuros pretende-se construir um sistema gráfico de apoio à decisão na identificação de dados anômalos, com o melhor projeto de aplicação desenvolvido nesta pesquisa, a RBF-DDA, com o intuito de apontar dados de vazão considerados anômalos para auxiliar o trabalho de técnicos capacitados em consistência e utilizá-lo no controle de qualidade de postos hidrológicos sob gestão do Instituto Tecnológico SIMEPAR, bem como em projetos de consistência em outras bacias do país.

Para projetos científicos futuros, serão estudadas alternativas para expressão da probabilidade posterior com outras representações para o Teorema de Bayes. Além disso, almeja-se expandir e analisar o desempenho dos métodos estudados para outras variáveis, por exemplo, diretamente para os dados de cota, e também para dados de precipitação. Também serão investigados o grande número de falsos alertas gerados pelos métodos bem como maneiras de minimizá-los. E, por fim, pretende-se estudar, e desenvolver projetos de aplicações para outros métodos de detecção de anomalias aplicados ao controle de qualidade de dados hidrológicos.

REFERÊNCIAS

ANA. **Sistema de Informações Hidrológicas**. 2013. Disponível em: http://www2.ana.gov.br/Paginas/servicos/informacoeshidrologicas/redehidro.aspx, acesso em: Dez 2013.

BERGMEIR, C.; BENÍTEZ, J. M. Neural networks in R using the stuttgart neural network simulator: RSNNS. **Journal of Statistical Software**, v. 46, n. 7, p. 1–26, 2012. Disponível em: $<$http://www.jstatsoft.org/v46/i07/$>$.

BERTHOLD, M. R.; DIAMOND, J. Boosting the performance of rbf networks with dynamic decay adjustment. **Advances in Neural Information Processing**, v. 7, p. 521–528, 1995.

BREDA, Â. Padrões de inconsistências no monitoramento automático de cota fluviométrica. In: **XX Simpósio Brasileiro de Recursos Hídricos**. Bento Gonçalves, Rio Grande do Sul, Brasil: [s.n.], 2013.

BREDA, Â.; NEGRÃO, A. C. **Relatório da Análise de Consistência dos Dados Hidrológicos na Bacia do Rio Iguaçu: Métodos e Resultados**. Curitiba: SIMEPAR, 2012.

CASTRO, L. N. **Fundamentals of Natural Computing: Basic Concepts, Algorithms, and Applications**. New York: Taylor e Francis, 2006.

CAVAZOS, T. Using self–organizing maps to investigate extreme climate events: An application to wintertime precipitation in the balkans. **Journal of Climate**, v. 13, p. 1718–1732, 1999.

CHEN, C.-S.; HUANG, H.-C. An improved c_p criterion for spline smoothing. **Journal of Statistical Planning and Inference**, v. 141, p. 445–452, 2010.

FOX, J. **An R and SPlus Companion to Applied Regression**. California: SAGE Publications, 2002.

FREIRE, L. S. **Uso de Rede Neural na Obtenção de Previsão Hidrológica Probabilística**. Trabalho de Conclusão de Curso — Universidade Federal do Paraná, Curitiba, 2009.

HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. 2. ed. Ontario: Prentice Hall, 1998.

HUANG, C. *et al.* On using smoothing spline and residual correction to fuse rain gauge observations and remote sensing data. **Journal of Hydrology**, v. 508, p. 410–417, 2013.

HUDAK, M. J. Rce classifiers: Theory and practice. **Cybernetics and Systems: An International Journal**, v. 23, n. 5, p. 483–515, 1992.

JICA. **The master plan study on the utilization of water resources in Paraná state in the Federative Republic of Brazil:Final Report**. [S.l.]: Japan International Cooperation Agency, 1995a.

JOLLIFFE, I. T.; STEPHENSON, D. B. **Forecast Verification: A Practitioner's Guide in Atmospheric Science**. Chichester: Wiley, 2003.

KINDELAN, M.; BAYONA, V. Application of the rbf meshless method to laminar flame propagation. **Engineering Analysis with Boundary Elements**, v. 37, p. 1617–1624, 2013.

KOHONEN, T. **Self-Organizing Maps**. Berlin: Springer, 2001.

KOUIBIAA, A.; PASADAS, M. Approximation by smoothing variational vector splines for noisy data. **Journal of Computational and Applied Mathematics**, v. 211, p. 213–222, 2008.

LEITE, E. A. **Gestão do Valor da Informação Hidrometeorológica: O Caso dos Alertas de Inundação para Proteção de Bens Móveis em Edificações Residenciais de União da Vitória**. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2008.

LIU, H.; CHANDRASEKAR, V.; XU, G. An adaptive neural network scheme for radar rainfall estimation from wsr–88d observations. **Journal of Applied Meteorology**, v. 40, p. 2038–2050, 2001.

LUNDE, E. *et al.* **multic: Quantitative linkage analysis tools using the variance components approach**. [S.l.], 2013. R package version 0.3.8. Disponível em: <http://CRAN.R-project.org/package=multic>.

MONTGOMERY, E.; JR, O. L. **Redes Neurais: Fundamentos e Aplicações com Programas em C**. Rio de Janeiro: Ciência Moderna, 2007.

NEGRÃO, A. C. **Consistência de Dados Hidrológicos das Sub–bacias do Rio Iguacu**. Curitiba: SIMEPAR, 2011.

OLIVEIRA, A. L. I. de. **Neural Networks Forecasting and Classification-Based Techniques for Novelty Detection in Time Series**. Tese (Doutorado) — Universidade Federal de Pernambuco, Recife, 2004.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Curvas roc para avaliação de classificadores. **IEEE Latin America Transactions**, v. 6, p. 215–222, 2008.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2012. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>.

RUGGIERO, M. A. G.; LOPES, V. L. da R. **Cálculo Numérico: Aspectos Teóricos e Computacionais**. São Paulo: Pearson, 1996.

SING, T. *et al.* **ROCR: Visualizing the performance of scoring classifiers.** [S.l.], 2009. R package version 1.0-4. Disponível em: <http://CRAN.R-project.org/package=ROCR>.

SIQUEIRA, P. H. **Uma Nova Abordagem na Resolução do Problema do Caixeiro Viajante**. Tese (Doutorado) — Universidade Federal do Paraná, Curitiba, 2005.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. 4. ed. [S.l.]: Elsevier Science, 2009.

VALENÇA, M. **Aplicando Redes Neurais: Um Guia Completo**. Pernambuco: Livro Rápido, 2005.

VESANTO, J.; ALHONIEMI, E. Clustering of the self–organizing map. **IEEE Transactions on Neural Networks**, v. 11, p. 586–600, 2000.

WEHRENS, R.; BUYDENS, L. M. C. Self- and super-organising maps in r: the kohonen package. **J. Stat. Softw.**, v. 21, n. 5, 2007. Disponível em: <http://www.jstatsoft.org/v21/i05>.

XIANHAI, G. Study of emotion recognition based on electrocardiogram and rbf neural network. **Procedia Engineering**, v. 15, p. 2408–2412, 2011.

ZELL, A. *et al.* **SNNS: Stuttgart Neural Network Simulator**. 4.2. ed. University of Stuttgart and University of Tübingen, 1998. Disponível em: <http://www.ra.cs.uni-tuebingen.de/SNNS/>.

ANEXO A – ESCOLHA DOS PARÂMETROS

A.1 PARÂMETROS DE SUAVIDADE PARA *SMOOTH SPLINE*

TABELA 3: Valores de AUC para a escolha dos parâmetros de suavidade

Parâmetros \ Postos Hidrológicos	Porto Amazonas	União da Vitória
0.01	0.851	0.827
0.05	0.851	0.827
0.10	0.85	0.826
0.20	0.857	0.827
0.25	0.856	0.834
0.50	**0.867**	0.839
0.75	0.857	**0.848**
0.95	0.85	0.845

A.2 VARIAÇÃO DO PARÂMETRO θ^-

TABELA 4: Valores de AUC para a escolha do parâmetro θ^-

Parâmetros \ Postos Hidrológicos	Porto Amazonas	União da Vitória
1×10^{-1}	0.9248296	0.9400361
5×10^{-2}	0.9340437	0.9485482
1×10^{-2}	**0.9401079**	0.9610013
5×10^{-3}	0.9388251	0.9677903
1×10^{-3}	0.9347836	0.9720367
5×10^{-4}	-	0.9728914
1×10^{-4}	-	0.9738433
5×10^{-5}	-	**0.9738533**
1×10^{-5}	-	0.9711797
5×10^{-6}	-	0.9709968
1×10^{-6}	-	0.9706167

APÊNDICE A – PROBABILIDADE TOTAL E REGRA DE BAYES

Se $A_i, i = 1,2,\ldots,M$, são M eventos tais que $\sum_{i=1}^{M} P(A_i) = 1$, então a probabilidade de um evento arbitrário B ocorrer é dado por:

$$P(B) = \sum_{i=1}^{M} P(B|A)P(A_i) \tag{27}$$

onde $P(B|A)$ indica a probabilidade condicional de B ocorrer assumindo que A também ocorreu, e é definido por:

$$P(B|A) = \frac{P(B,A)}{P(A)} \tag{28}$$

e $P(B,A)$ é a probabilidade conjunta dos dois eventos. A Equação 27 é conhecida como Teorema da Probabilidade Total. A partir da definição 28 a regra de Bayes implica em:

$$P(B|A)P(A) = P(A|B)P(B) \tag{29}$$

Isto pode ser estendido para variáveis aleatórias ou vetores descritos por funções de densidade de probabilidade e tem-se:

$$p(x|A)P(A) = P(A|x)p(x) \tag{30}$$

e:

$$p(x|y)p(y) = p(y|x)p(x) \tag{31}$$

e finalmente:

$$p(x) = \sum_{i=1}^{M} p(x|A_i)P(A_i) \tag{32}$$